Certificate course on AI and ML

Data Visualization

Prof. Kusum Deep
Full Professor (HAG), Department of Mathematics
Joint Faculty, MF School of Data Science and Artificial Intelligence
Indian Institute of Technology Roorkee, Roorkee – 247667

kusum.deep@ma.iitr.ac.in, kusumdeep@gmail.com

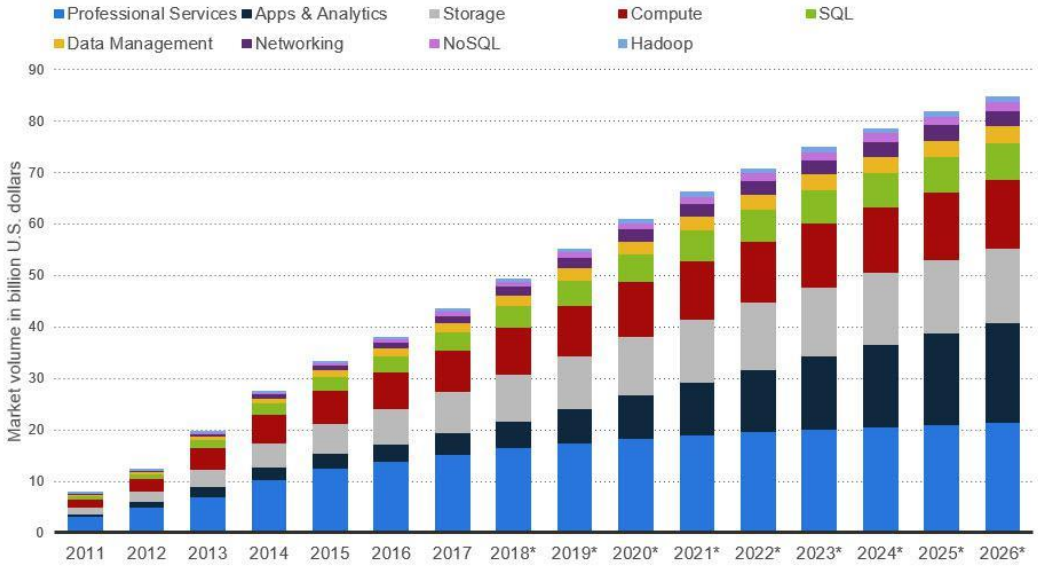
Today's Age of Big data

- Today, the amount of data being generated and collected has reached unprecedented levels, and has become a critical resource for businesses, governments, and other organizations.
- Because of increased usage of digital technologies like social media, mobile devices, IoT, machine learning algorithms, etc., data is being generated at an exponential rate. It is a challenge to extract relavant information from this huge amount of data.

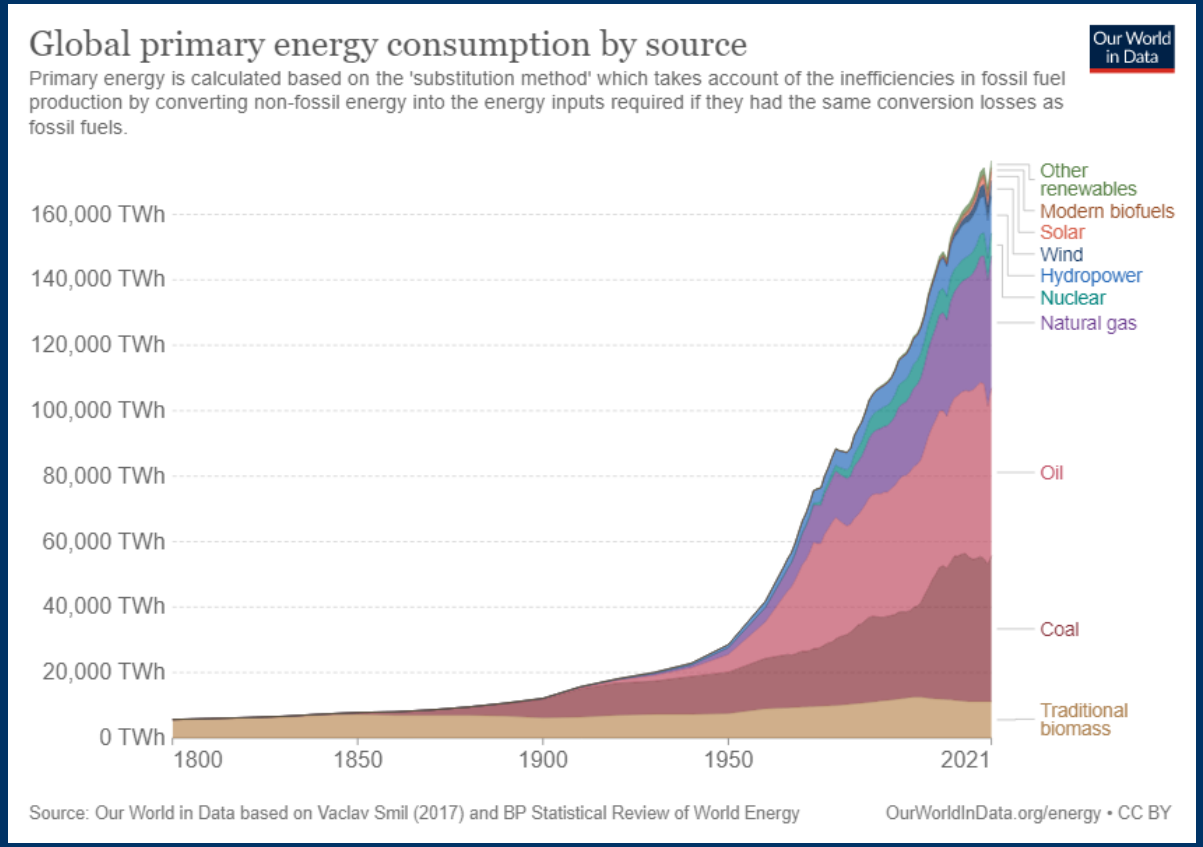
- This has brought about a number of significant changes, including the need for new technologies and approaches to manage and analyze data, the emergence of new job roles such as data scientists and data analysts, and the potential for data-driven insights to inform decision-making across a range of industries.
- At the same time, the age of big data has also raised concerns around data privacy, security, and ethical use, as organizations seek to balance the potential benefits of data with the need to protect individual rights and prevent unintended consequences.

Big Data Market Worldwide Segment Revenue Forecast 2011-2026

Big Data Market Forecast Worldwide from 2011 to 2026, by segment (in billion U.S. dollars)



statista


Global primary energy consumption by source
Primary energy is calculated based on the 'substitution method' which takes account of the inefficiencies in fossil fuel production by converting non-fossil energy into the energy inputs required if they had the same conversion losses as fossil fuels.
Our World in Data
160,000 TWh
140,000 TWh
120,000 TWh
100,000 TWh
80,000 TWh
60,000 TWh
40,000 TWh
20,000 TWh
0 TWh
1800
1850
1900
1950
2021
Other renewables
Modern biofuels
Solar
Wind
Hydropower
Nuclear
Natural gas
Oil
Coal
Traditional biomass
Source: Our World in Data based on Vaclav Smil (2017) and BP Statistical Review of World Energy
OurWorldInData.org/energy • CC BY

Market size of Big Data

Global big data market size is expected to grow from USD 138.9 billion in 2020 to USD 229.4 billion by 2025, at a compound annual growth rate (CAGR) of 10.6% during the forecast period.
It includes software, hardware, and services, as well as various industry verticals, including banking, healthcare, retail, and government. This is driven by the increasing volume and complexity of data, the need for organizations to gain insights from this data, and the availability of new technologies and tools for managing and analyzing data.

- COVID-19 also accelerated the adoption of big data technologies and solutions, as organizations seek to better understand and respond to the changing business landscape. This has further fueled the growth of the big data market, particularly in healthcare, supply chain management, and remote work technologies

Challenges & Opportunities in Big Data

Challenges:

Volume: is the biggest challenge. With the growth of the internet and connected devices, the amount of data being generated is increasing exponentially.

Velocity: Speed is a challenge. data is being generated in real-time, and businesses need to be able to process and analyze it quickly to gain insights and make informed decisions.

Variety: Data comes in many different forms, including structured and unstructured data, text, images, and videos. The challenge is in terms of managing and analyzing the data.

Veracity: The quality of data is a challenge. Data can be incomplete, inconsistent, or inaccurate, which can lead to incorrect insights and decisions.

Security and Privacy: Businesses need to ensure that they are collecting, storing, and using data in a secure and ethical manner.

Opportunities:

Improved Decision Making: Big Data provides businesses with insights that can help them make informed decisions. By analyzing large amounts of data, businesses can identify patterns, trends, and relationships that they may have otherwise missed.

Increased Efficiency: Big Data can help businesses automate processes and improve efficiencies. By analyzing data, businesses can identify areas for improvement and optimize their operations.

Better Customer Experience: By analyzing customer data, businesses can gain insights into customer behavior, preferences, and needs. This can help businesses personalize their offerings and improve the overall customer experience.

New Revenue Streams: Big Data can help businesses identify new revenue streams by analyzing customer behavior, market trends, and other data sources. This can lead to the development of new products and services.

Competitive Advantage: Businesses that are able to effectively manage and analyze Big Data can gain a competitive advantage by making better decisions, improving efficiencies and delivering better customer experiences.

Data Visualization & Big Data

Data visualization is the process of representing data in a graphical or pictorial format, which helps in understanding complex information quickly and easily. Ways in which data visualization can help in Big Data analytics are:

Identifying patterns and trends: With data visualization, patterns and trends can be identified more easily.

Simplifying complex data: Users can quickly understand the relationships between different data points and identify insights that might have been missed in a more traditional analysis.

Exploring data in real-time: It tools can provide real-time data exploration and analysis, which can help businesses respond quickly to changing conditions and take immediate action based on the insights gained from the data.

Enhancing communication: It can help enhance communication by presenting data in a format that is easy to understand, which can help stakeholders across different departments or teams to collaborate more effectively and make informed decisions.

Enabling data-driven decision making: It can help in making data-driven decisions. By presenting data in a visual format, decision-makers can quickly and easily understand the data and identify the best course of action.

Benefits of Data Visualization

- Simplifies complex data
- Improves decision-making
- Enhances communication
- Saves time
- Improves data accuracy
- Increases engagement

Challenges in Data Visualization

- Data quality
- Data complexity
- Choosing the right visualization
- Interactivity
- User experience
- Accessibility

Considerations for Data Visualization

- Purpose
- Audience
- Data quality
- Design
- Clarity
- Context
- Interactivity

Data Visualization Pipeline

- Data acquisition
- Data cleaning
- Data exploration
- Data modeling
- Visualization design
- Visualization implementation
- Publication and distribution

Points to remember while Visualizating Data

Data Ink Ratio (DIR) is a concept introduced by Edward Tufte in his book "The Visual Display of Quantitative Information“. It refers to the proportion of ink on a page that is used to represent actual data. The purpose of DIR is to minimize non-data ink in a visualization, which can detract from the clarity and effectiveness of the visualization.

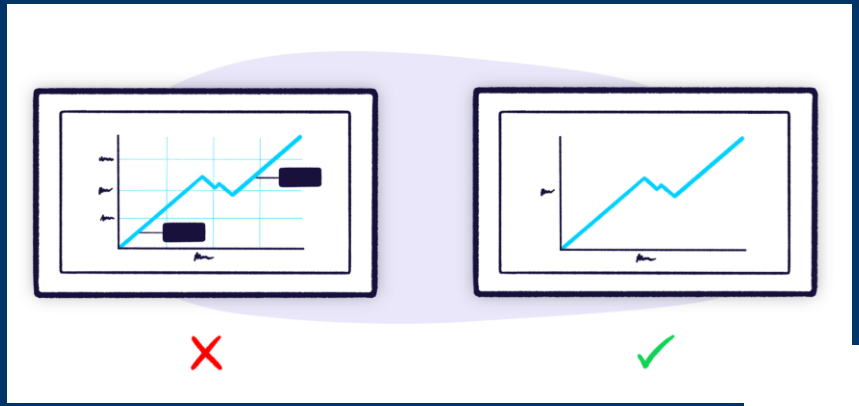
It is calculated by dividing the amount of ink used to represent the data by the total amount of ink used in the visualization.

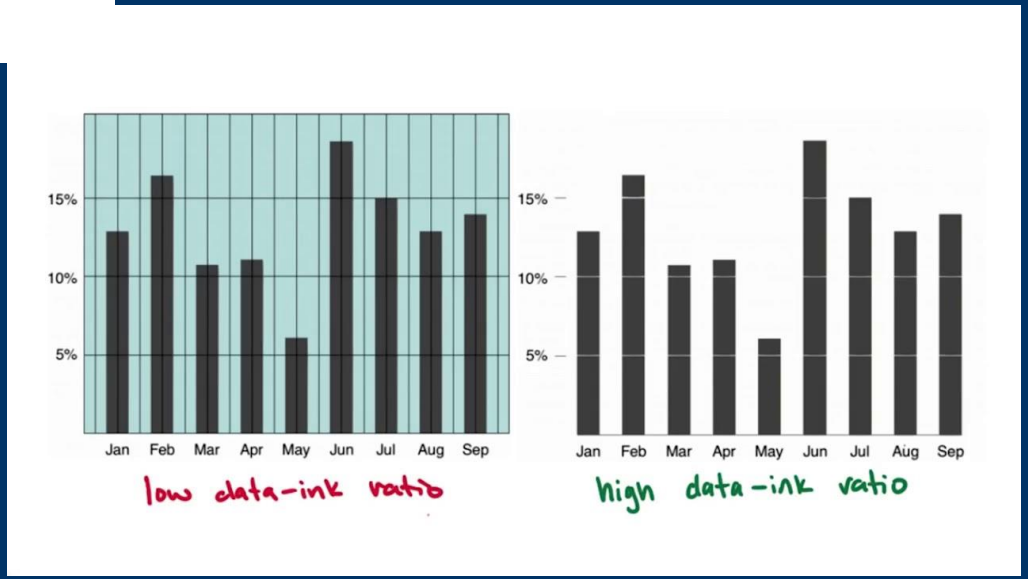
The lesser the DIR, the better the visualization. (DIR <=1)

$$\text{Data} - \text{Ink Ratio} = \frac{\text{Data} - \text{ink}}{\text{Total ink used to print the graphics}}$$

$=$ proportion of graphic's ink devoted to the nonredudant display of data information

$= 1 -$ proportion of a graphic that can be erased without loss of data information




15%
10%
5%
Jan Feb Mar Apr May Jun Jul Aug Sep
low data-ink ratio
15%
10%
5%
Jan Feb Mar Apr May Jun Jul Aug Sep
high data-ink ratio

THE DO'S AND DON'TS OF CHART MAKING

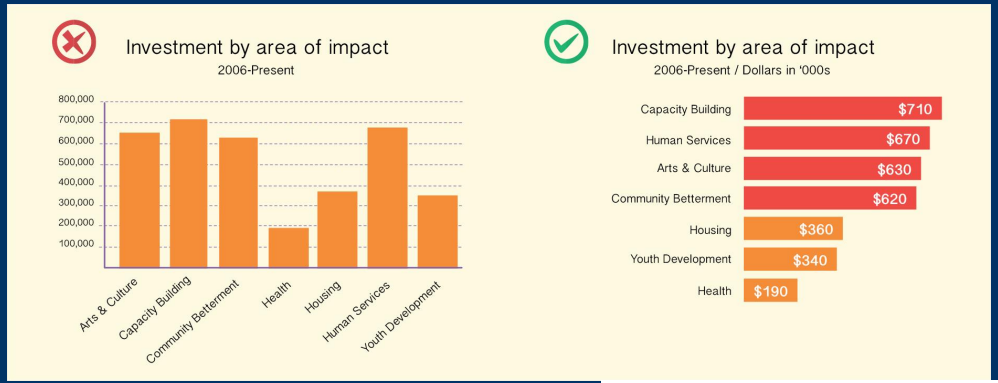
DO

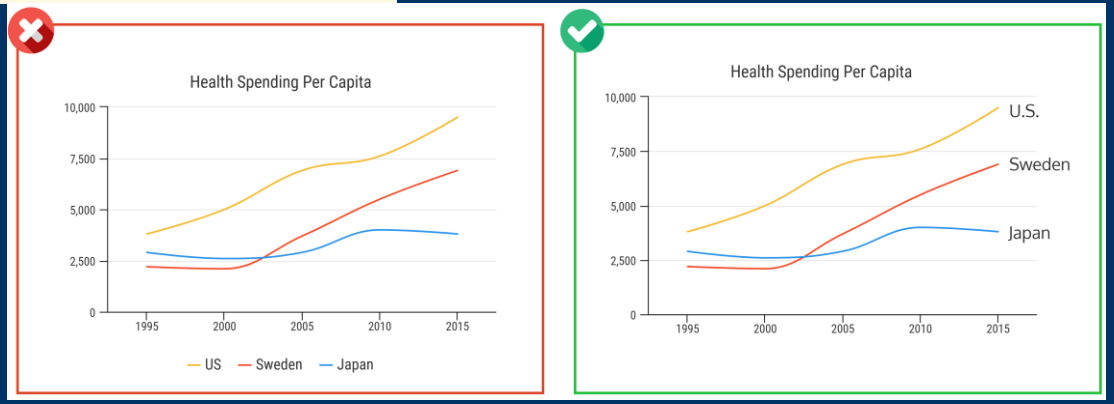
- Use appropriate charts, including horizontal bar graphs
- Use the full axis
- Keep it simple, especially with animations, and make sure with a squint test
- Use color to contrast and highlight data
- Ask others for opinions

DON'T

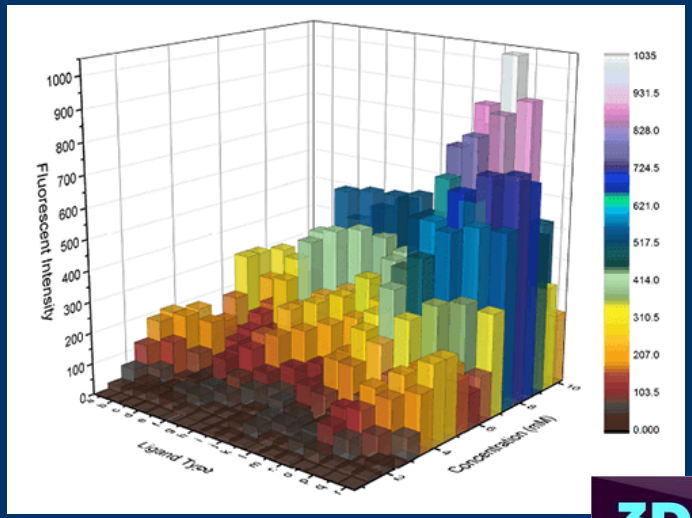
- Change chart styles partway through a comparison
- Overload charts with unimportant data, more than six colors, or too many animations
- Use a pie chart, especially one with more than seven wedges
- Use combinations with similar colors (red/orange and green, blues and greens)
- Sacrifice important data

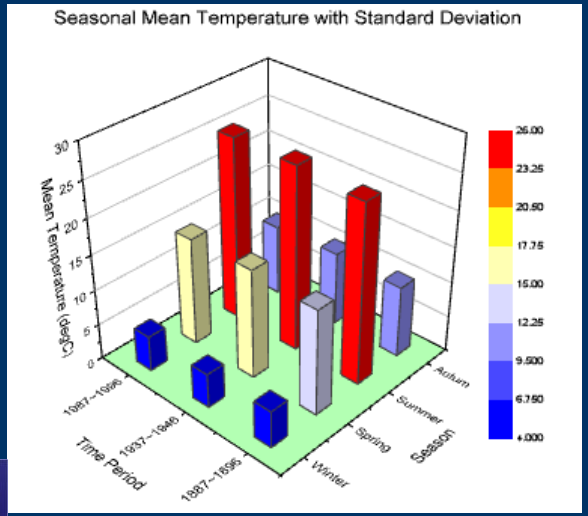
What makes a good chart?





3-D Charts

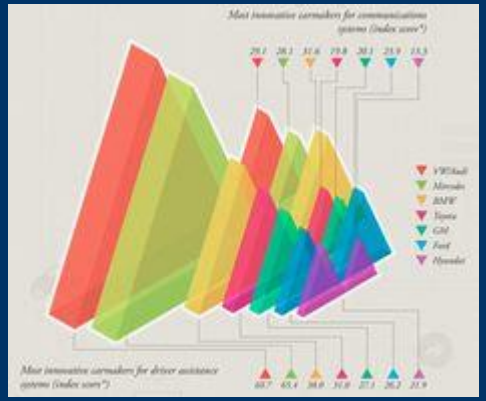

Fluorescent Intensity
Ligand Type
Concentration (mM)
1035
931.5
828.0
724.5
621.0
517.5
414.0
310.5
207.0
103.5
0.000

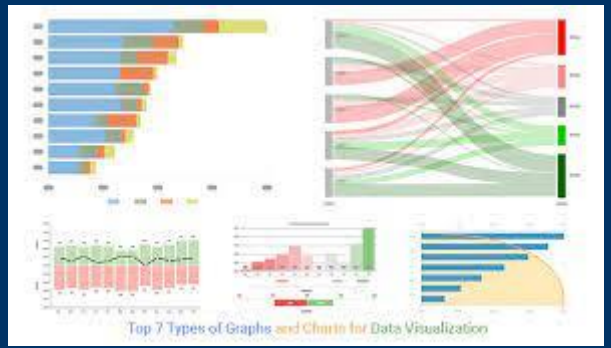

Seasonal Mean Temperature with Standard Deviation
Mean Temperature (deg C)
Time Period
Season
Autumn
Summer
Spring
Winter
26.00
23.25
20.50
17.75
15.00
12.25
9.500
6.750
4.000


3D
Bar
Chart
You
Haven't
Seen
This
Before
STEP A
STEP B
STEP C
Creative Venus

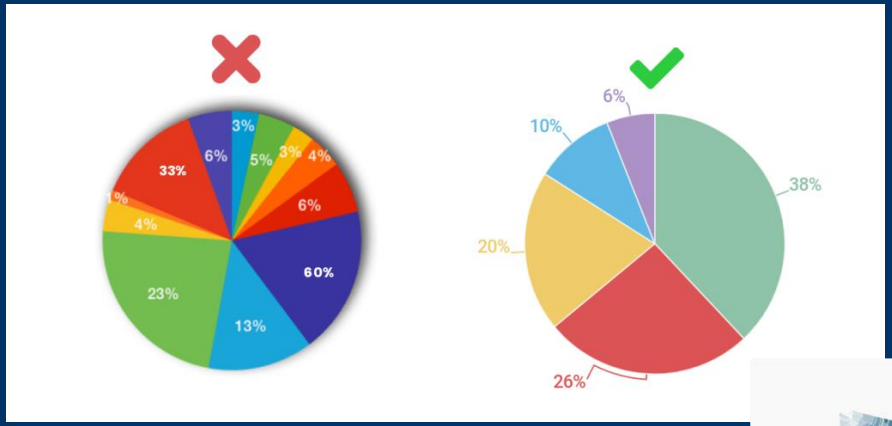
3-D charts

Avoid Chartjunk

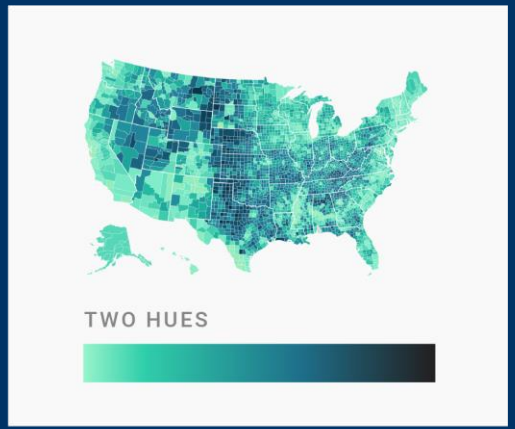




Color scales



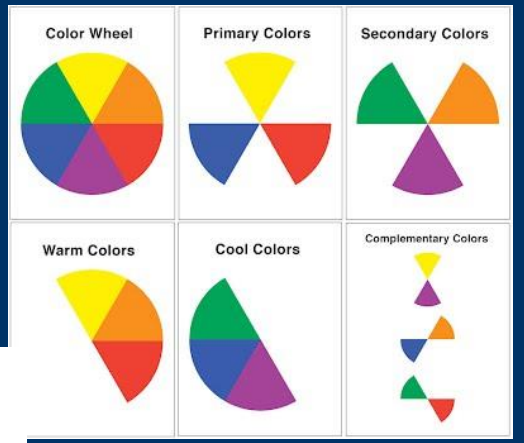
NOT SO BAD

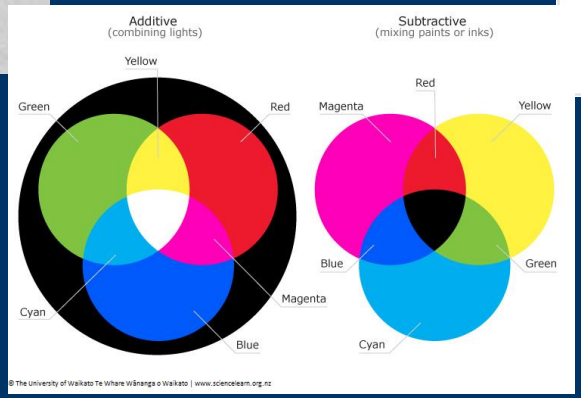


EVEN BETTER

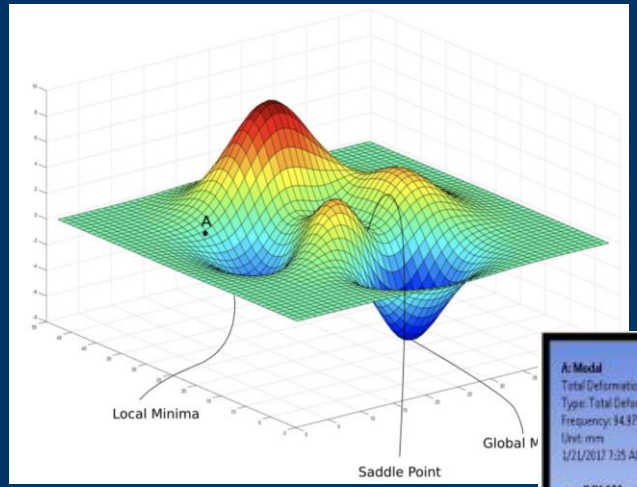

Top 3 per Location with Rating
Wrong
Above Three Rating
No
Yes
Location
Product
Barista
Cake
Latte
Latte
Coffee
Belfrance Chocolate
Coffee
Dark Chocolate
Milk Chocolate
Toffee
Brats and Cuties
Arts and Crafts
Playtime
Day Care
Eat at Joes
Pasta
Fries
Fries
Hamburger
Soda
Beer
This is correct
This is WRONG
Latte is split based on Above Three Rating. It should be single bar with two color, just like Barista, Cake

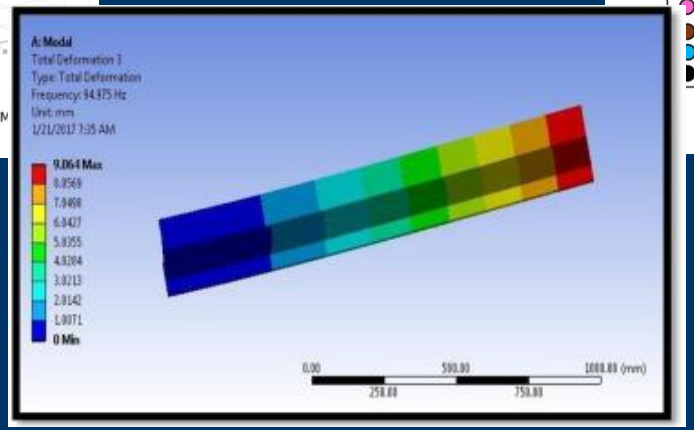
Colour meaning charts

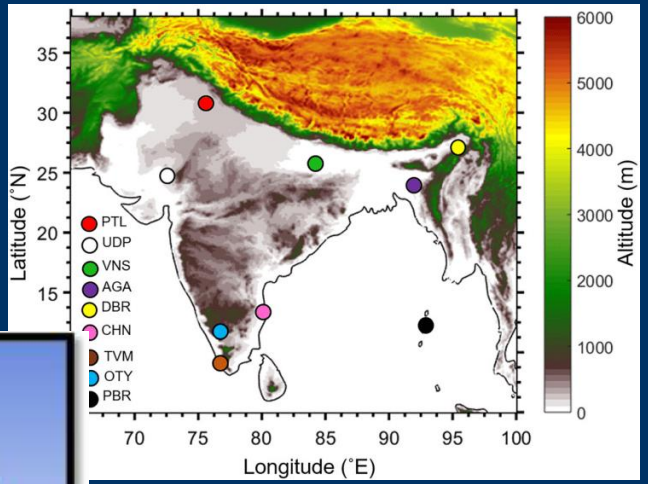




Colour grading







Data Visualization Principles

- Purpose
- Clarity
- Accuracy
- Context
- Simplicity
- Consistency
- Color
- Interactivity
- Storytelling

Data Visualization Principles

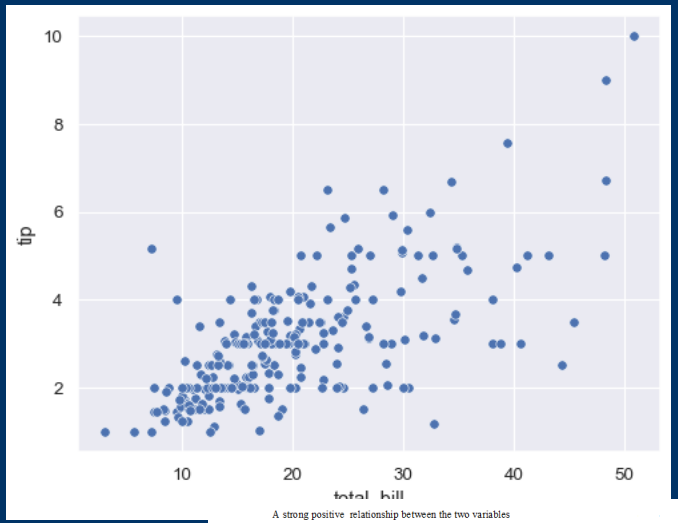
- Improve vision
  1. Reduced clutter, Make data stand out
  2. Use visually prominent graphical elements
  3. Use proper scale lines and a data rectangle
  4. Reference lines, labels, notes, and keys
  5. Superposed data set

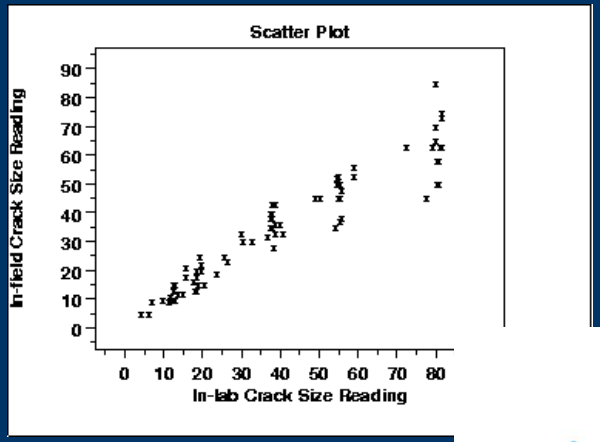
- Improve understanding
  1. Provide explanations and draw conclusions
  2. Use all available space
  3. Align juxtaposed plots
  4. Use log scales when appropriate

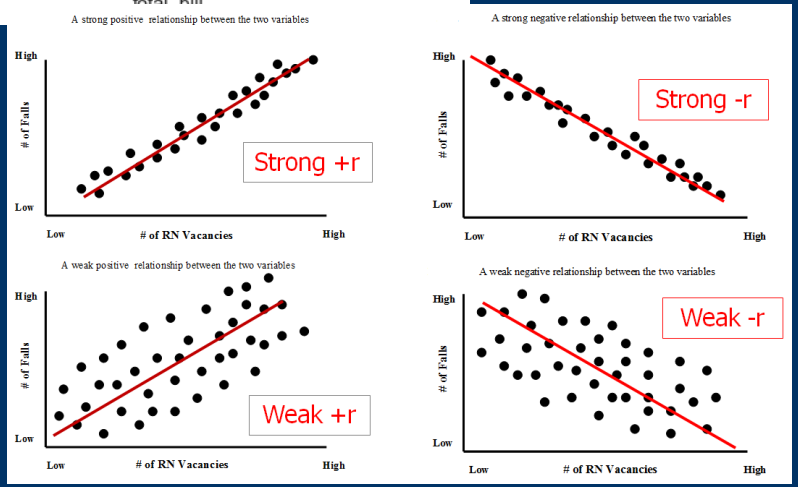
Some Common Data Visualization Plots

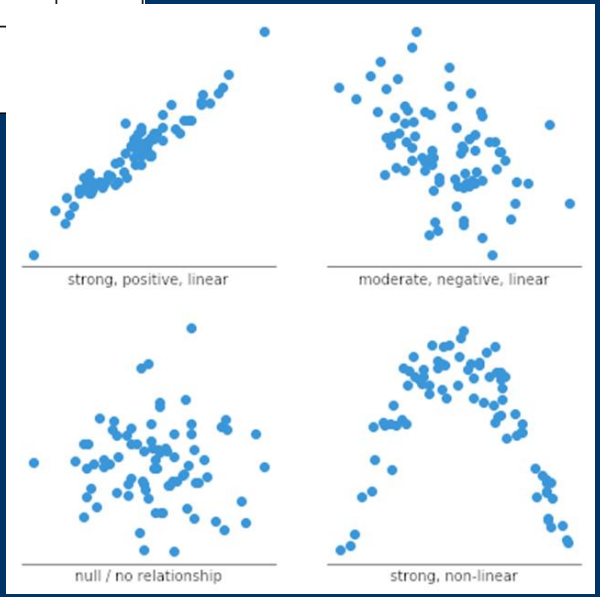
Scatterplot

A scatter plot displays the relationship between two quantitative variables. It is a collection of data points, where each point represents a value for both variables being compared, plotted on a two-dimensional coordinate plane. Scatter plots are commonly used to identify patterns or relationships between the two variables and to identify any outliers or unusual data points. They can also be used to visually compare data from two or more groups or populations.

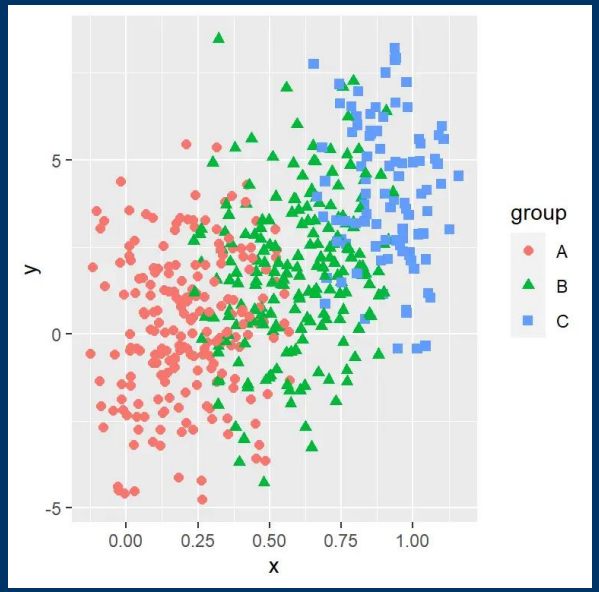

tip
total_bill

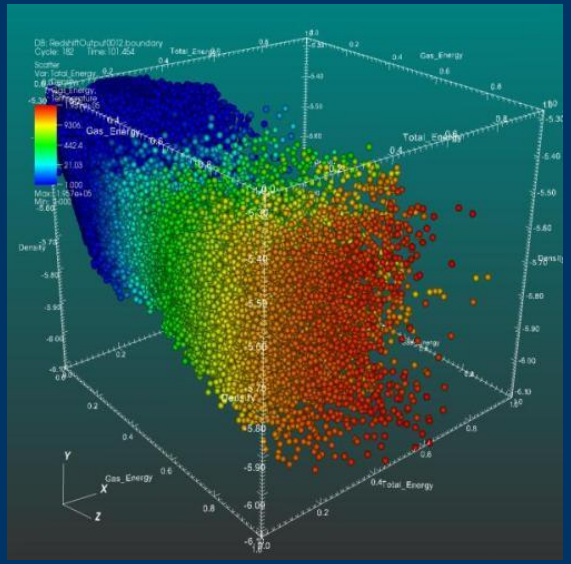

Scatter Plot
In-field Crack Size Reading
In-lab Crack Size Reading

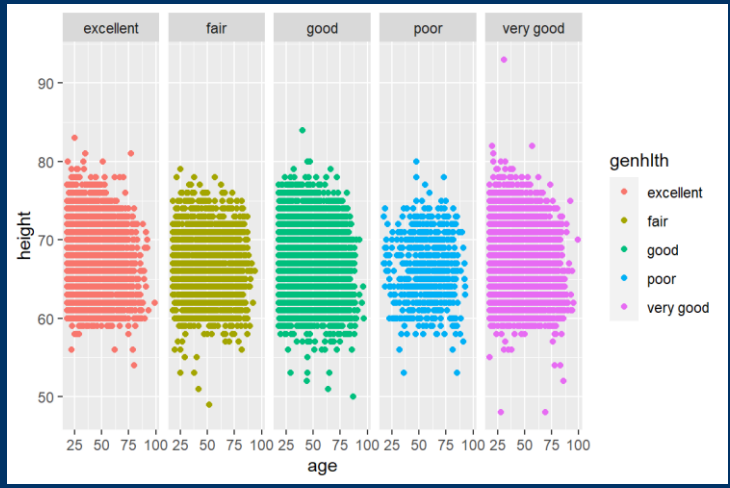

A strong positive relationship between the two variables
Strong +r
A strong negative relationship between the two variables
Strong -r
A weak positive relationship between the two variables
Weak +r
A weak negative relationship between the two variables
Weak -r
# of Falls
# of RN Vacancies
High
Low


strong, positive, linear
moderate, negative, linear
null / no relationship
strong, non-linear

Scatter plot for categorical variables





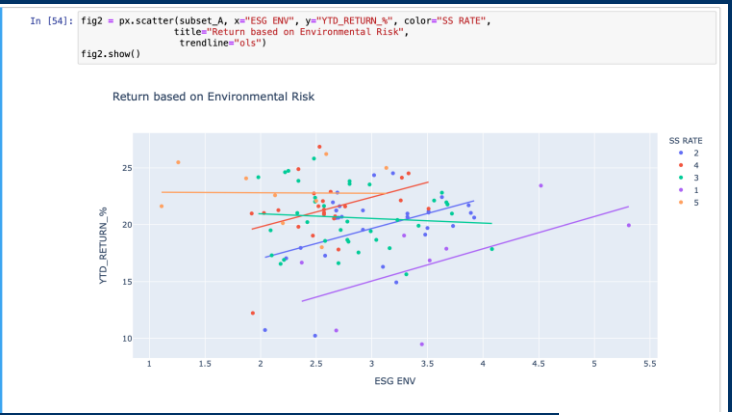


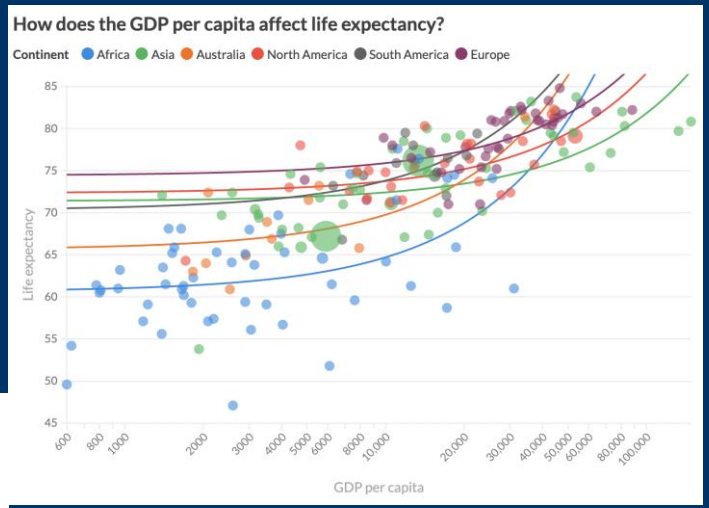
Ex: Create a scatter plot of categorical variables based on 3 variables, by assigning each category a different color or shape and plot each data point in a 3D coordinate system. Use this data:

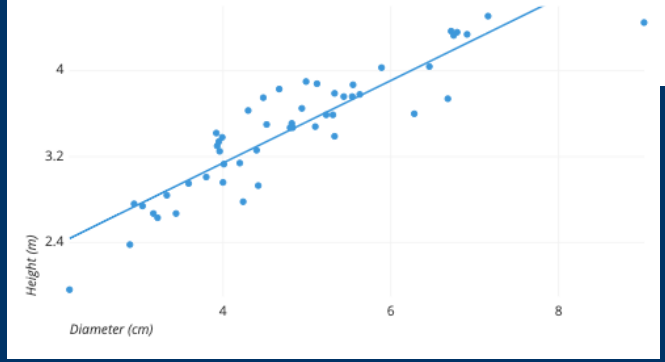
ategory	Variable 1	Variable 2	Variable 3
A	1.2	3.4	5.6
B	2.1	4.5	6.7
C	3.4	2.1	4.5
D	4.5	5.6	3.4
E	5.6	1.2	2.1
A	6.7	6.7	1.2
B	2.1	2.1	6.7
C	3.4	4.5	5.6
D	4.5	3.4	2.1
E	5.6	5.6	4.5
A	6.7	1.2	3.4
B	2.1	6.7	4.5
C	3.4	2.1	2.1
D	4.5	4.5	6.7
E	5.6	3.4	5.6
A	6.7	5.6	4.5
B	2.1	3.4	3.4
C	3.4	6.7	1.2
D	4.5	2.1	4.5
E	5.6	4.5	6.7
A	6.7	3.4	2.1
B	2.1	5.6	5.6
C	3.4	1.2	6.7
D	4.5	4.5	3.4
E	5.6	2.1	2.1

Scatter plot with trend lines

```
In [54]: fig2 = px.scatter(subset_A, x="ESG ENV", y="YTD_RETURN_%", color="SS RATE",
                           title="Return based on Environmental Risk",
                           trendline="ols")
         fig2.show()
```

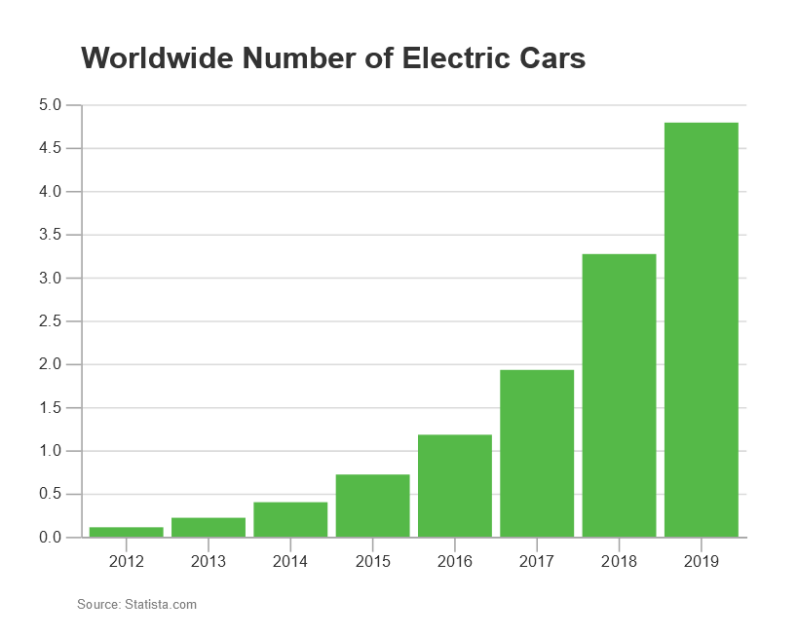


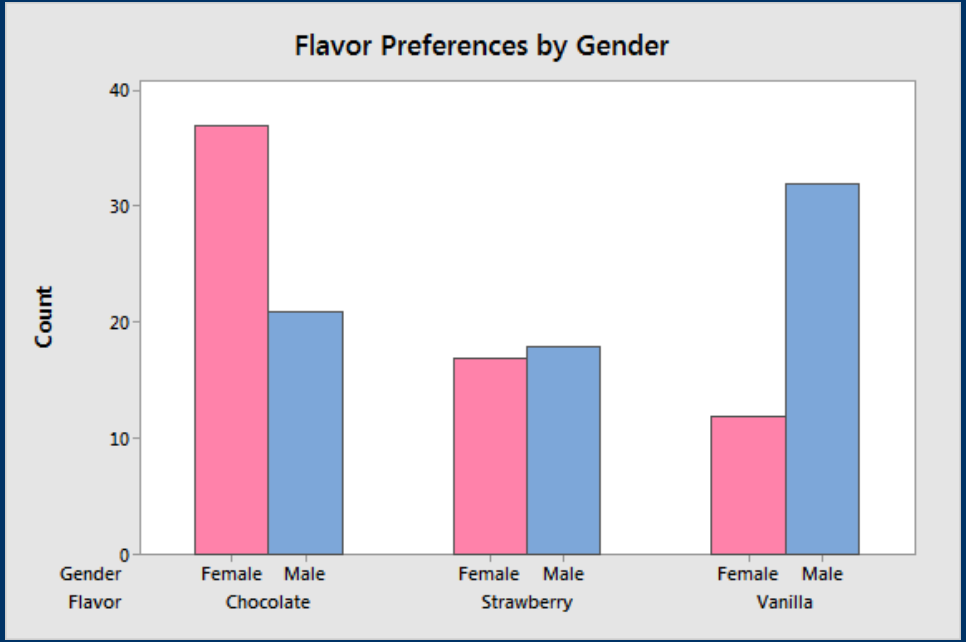


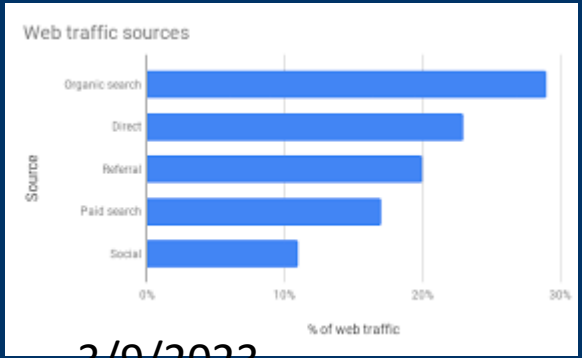


Bar plots

Bar plot displays categorical data with rectangular bars. The height or length of each bar corresponds to the frequency or proportion of data in each category. Bar plots are often used to compare the frequencies or proportions of different categories or groups. They can also be used to display changes in data over time or to highlight differences between subgroups of data. Bar plots can be created using a variety of software tools and can be customized with colors, labels, and other visual elements to enhance their effectiveness.


Worldwide Number of Electric Cars
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
2012
2013
2014
2015
2016
2017
2018
2019
Source: Statista.com


Flavor Preferences by Gender
Count
40
30
20
10
0
Gender
Flavor
Female
Male
Chocolate
Female
Male
Strawberry
Female
Male
Vanilla


Web traffic sources
Source
Organic search
Direct
Referral
Paid search
Social
0%
10%
20%
30%
% of web traffic

Ex: Suppose we conducted a survey of 50 people, asking them what their favorite fruit is. The results are as follows:

10 people chose apples

20 people chose bananas

5 people chose oranges

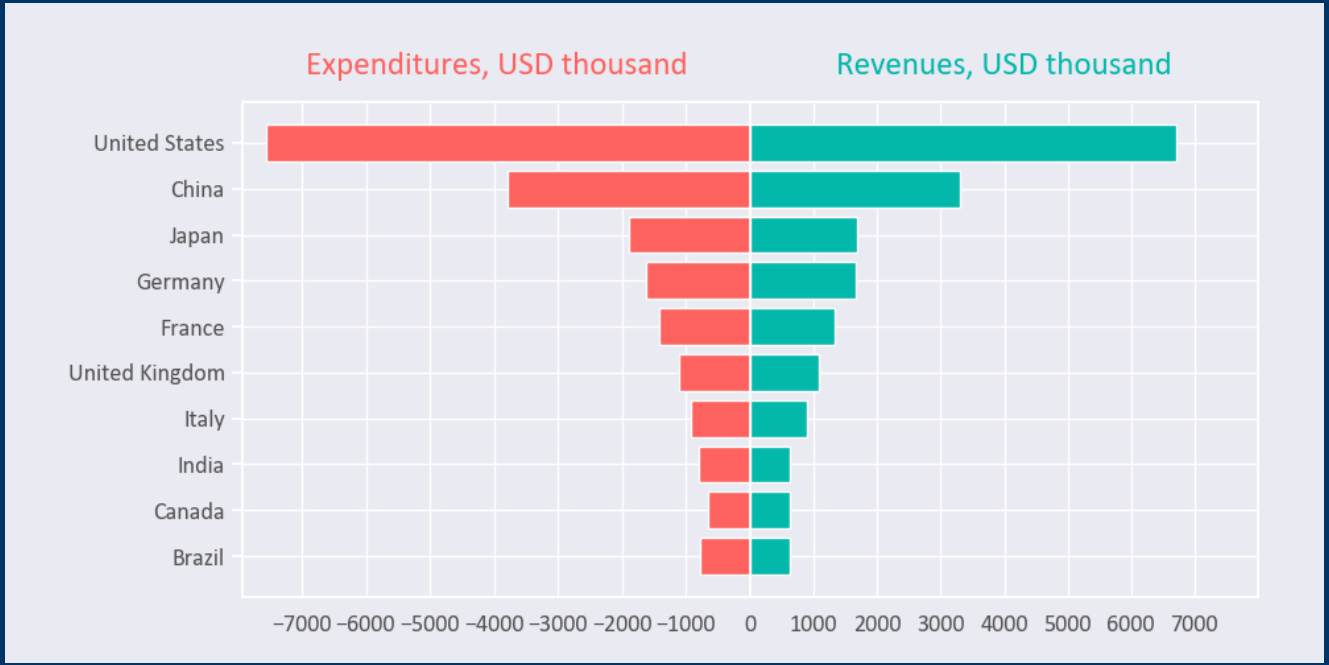
10 people chose strawberries

5 people chose pineapples

```
import matplotlib.pyplot as plt
fruits = ['apples', 'bananas', 'oranges', 'strawberries', 'pineapples']
counts = [10, 20, 5, 10, 5]
plt.bar(fruits, counts)
plt.title('Favorite Fruits Survey Results')
plt.xlabel('Fruit')
plt.ylabel('Count')
plt.show()
```

This will produce a bar plot with five bars, one for each fruit, with the height of each bar representing the number of people who chose that fruit as their favorite.

Bidirectional Bar plots



Ex: Suppose we have two groups, group A and group B, and we want to compare their scores on a test. The scores are as follows:

Group A: 75, 80, 85, 90, 95

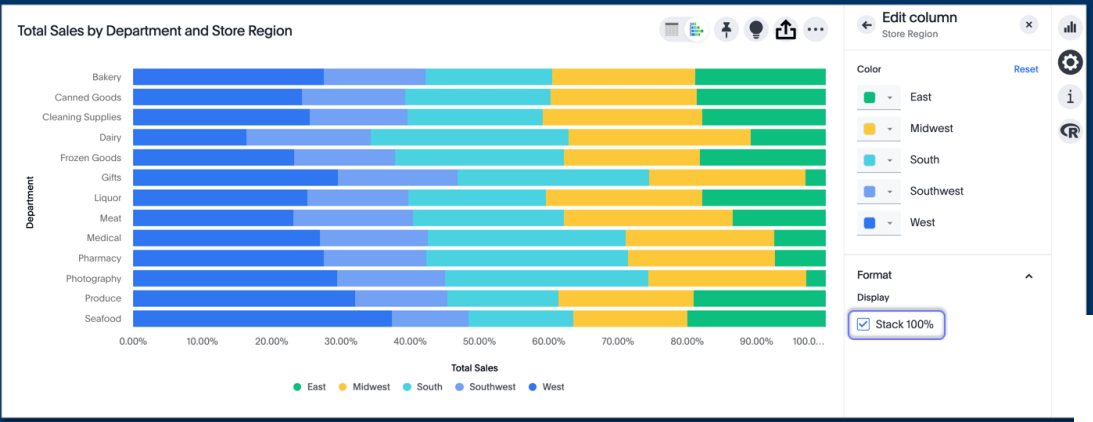
Group B: 65, 70, 75, 80, 85

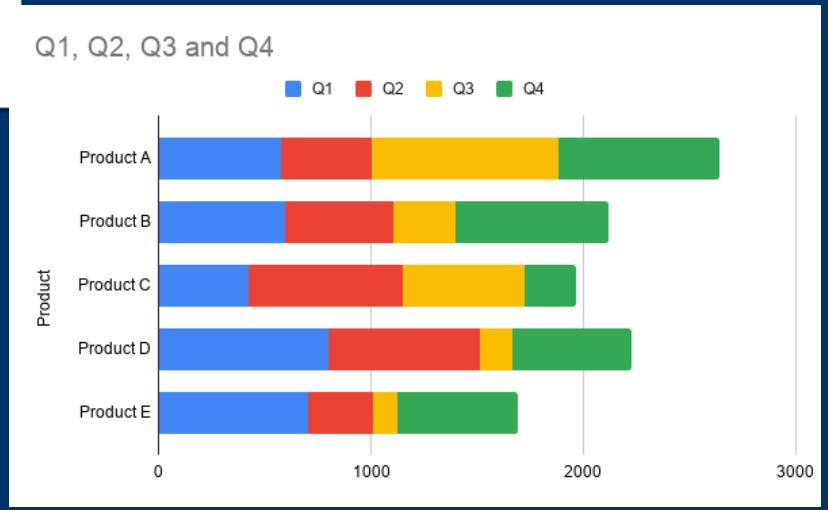
Create a bidirectional bar plot to visualize these results using python:

```
import matplotlib.pyplot as plt
import numpy as np
group_A = [75, 80, 85, 90, 95]
group_B = [65, 70, 75, 80, 85]
ind = np.arange(len(group_A))  # the x locations for the groups
width = 0.35  # the width of the bars
fig, ax = plt.subplots()
rects1 = ax.bar(ind - width/2, group_A, width, label='Group A')
rects2 = ax.bar(ind + width/2, group_B, width, label='Group B')
# Add some text for labels, title and custom x-axis tick labels, etc.
ax.set_ylabel('Scores')
ax.set_title('Scores by group')
ax.set_xticks(ind)
ax.set_xticklabels(['Test ' + str(i+1) for i in range(len(group_A))])
ax.legend()
# Add a horizontal line at y=0 to show the zero point
ax.axhline(y=0, color='gray', linewidth=1)
plt.show()
```

This will produce a bidirectional bar plot with two sets of bars, one for each group. The bars are positioned at each test number and the height of each bar represents the score on that test. The bars are colored differently for each group, and there is a legend to indicate which group each color represents. The y-axis is centered at zero, and there is a horizontal line at y=0 to help visualize the difference between the two groups.

Stacked Bar plots





Ex: Suppose we have a company that sells three products: Product A, Product B, and Product C. We want to visualize the total sales for each product broken down by region (North, South, East, West). The sales data is as follows:

	North	South	East	West
Product A	100	150	200	50
Product B	75	100	125	25
Product C	50	75	100	20

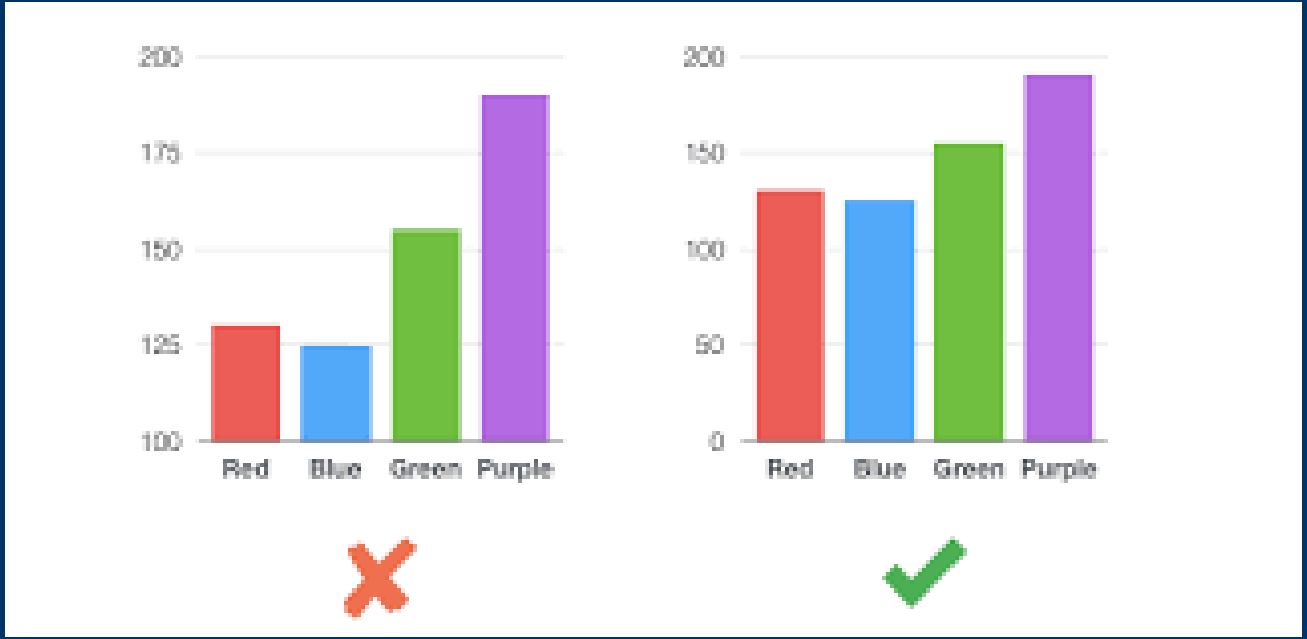
We can create a stacked bar plot to visualize these results as follows:

```
import matplotlib.pyplot as plt
import numpy as np
products = ['Product A', 'Product B', 'Product C']
regions = ['North', 'South', 'East', 'West']
sales = np.array([[100, 150, 200, 50],
                  [75, 100, 125, 25],
                  [50, 75, 100, 20]])
fig, ax = plt.subplots()
bottom = np.zeros(len(regions))
for i, product in enumerate(products):
    ax.bar(regions, sales[i], bottom=bottom, label=product)
    bottom += sales[i]
ax.set_xlabel('Region')
ax.set_ylabel('Sales')
ax.set_title('Total Sales by Product and Region')
ax.legend()
plt.show()
```

This will produce a stacked bar plot with three sets of bars, one for each product. Each set of bars is stacked on top of each other to show the total sales for that product broken down by region. The x-axis shows the four regions and the y-axis shows the total sales. Each set of bars is colored differently to represent the different products, and there is a legend to indicate which color represents which product.

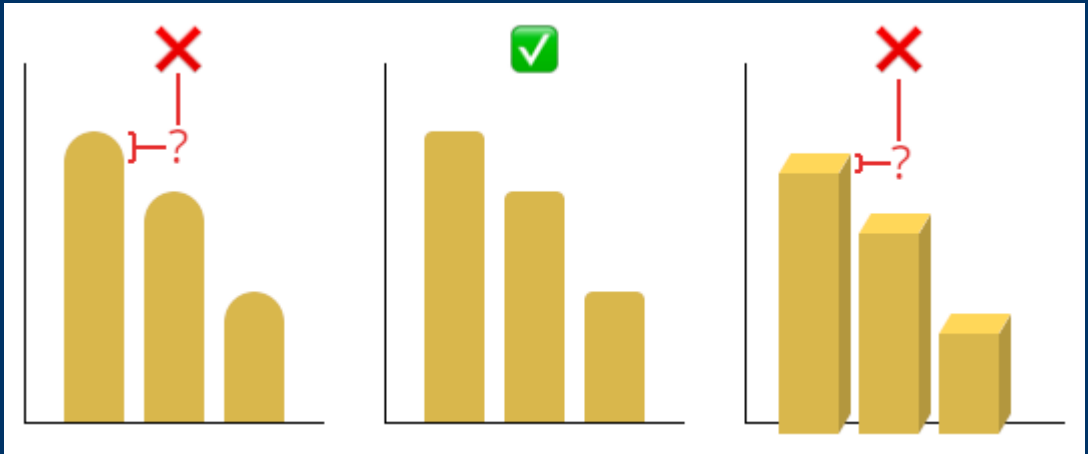
Bar Plots – Best Practices

Use a common zero-value baseline



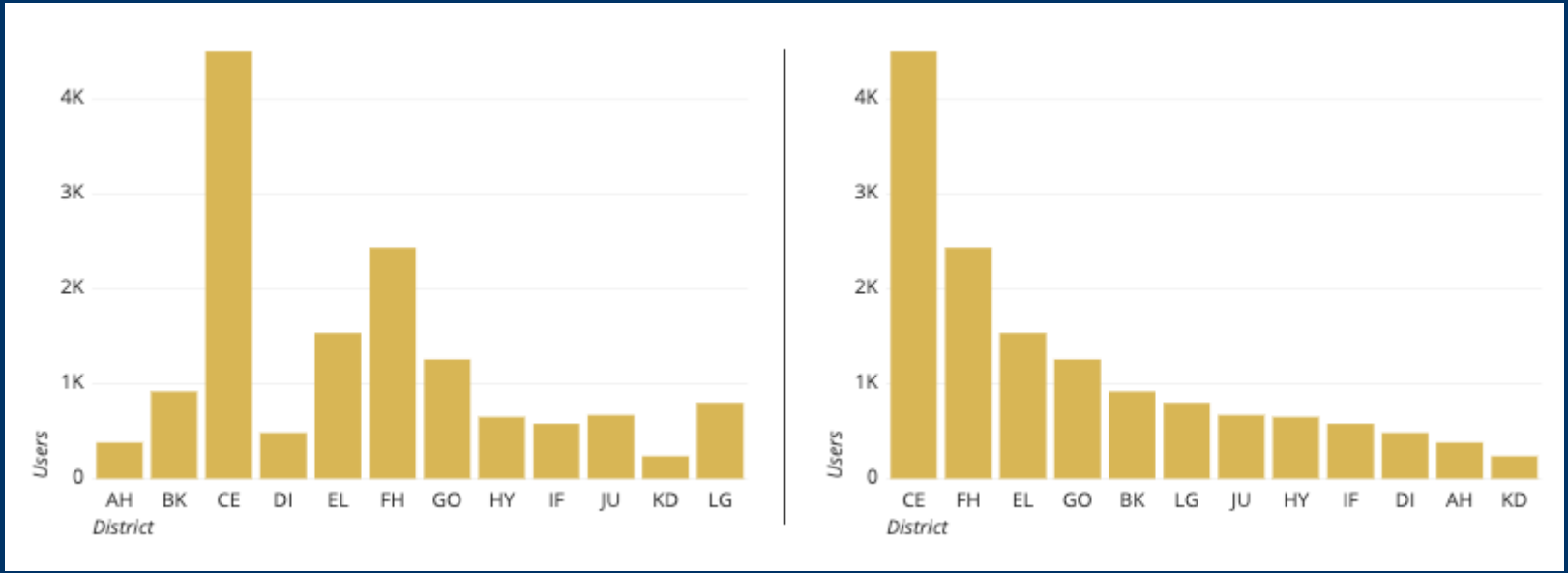
Bar Plots – Best Practices

Maintain rectangular forms for your bars

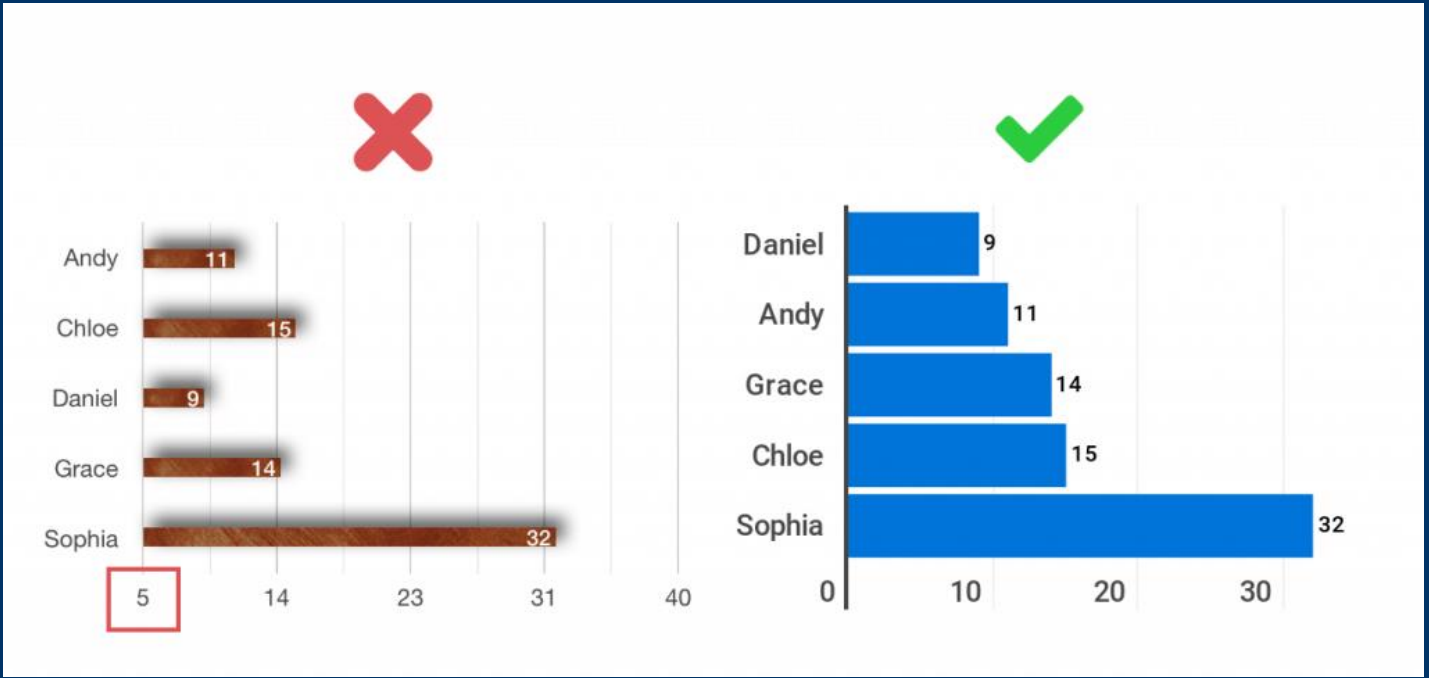


Bar Plots – Best Practices

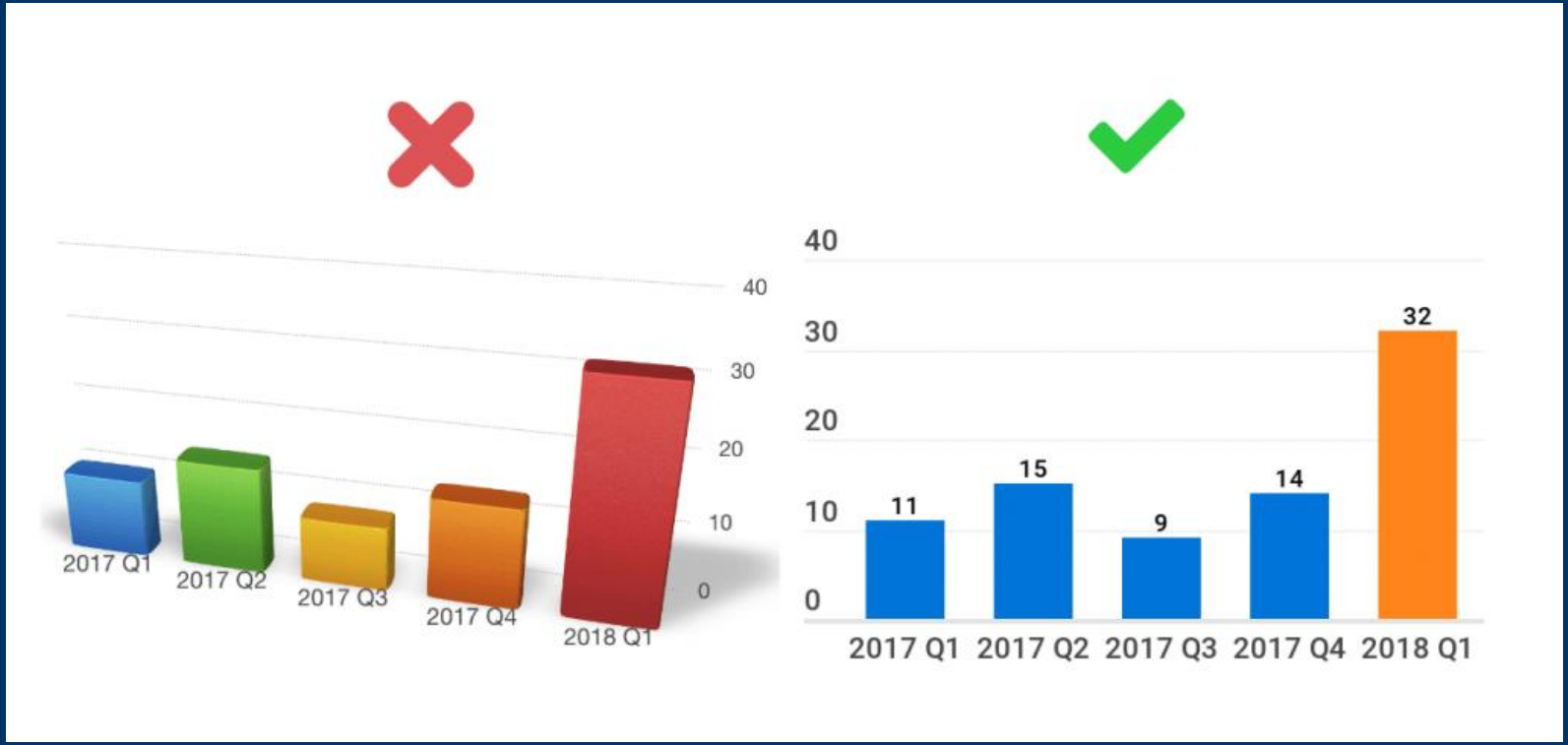
Consider the ordering of category levels



Bar Plots – Best Practices

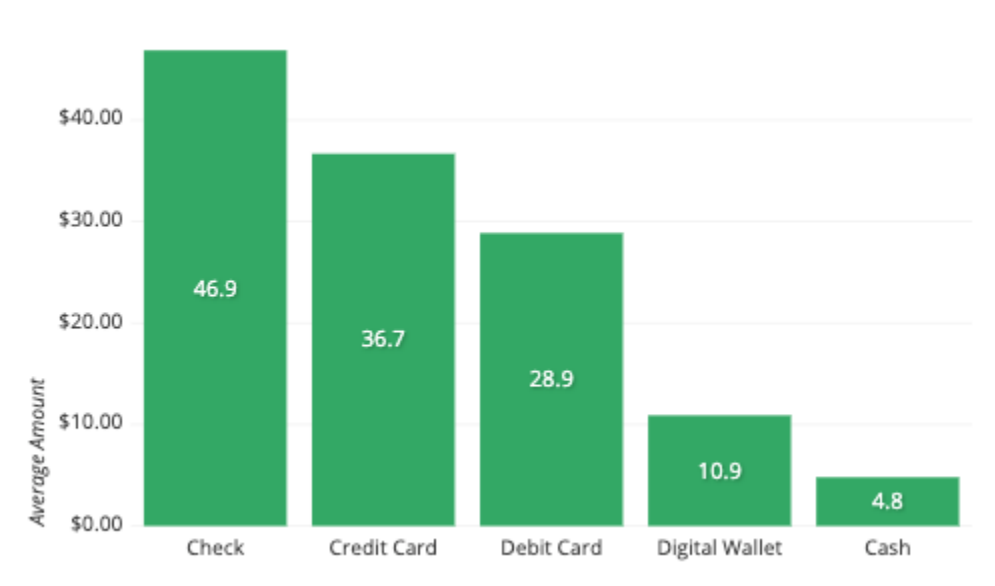


Bar Plots – Best Practices



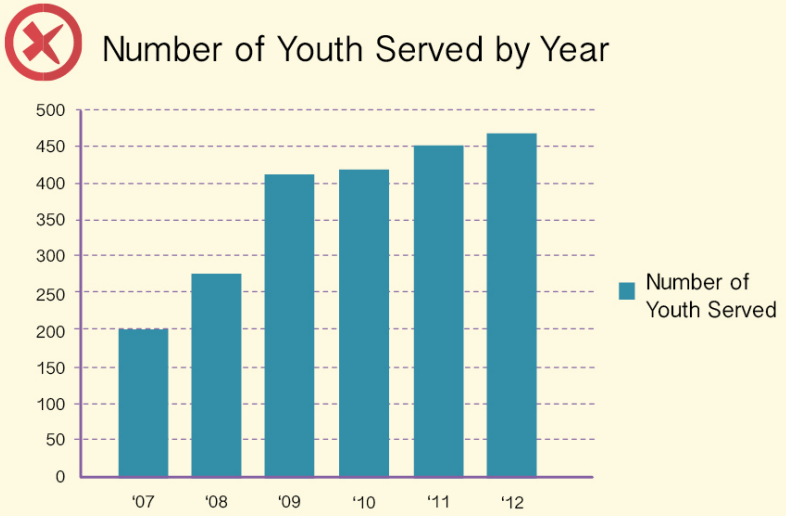
Bar Plots – Best Practices

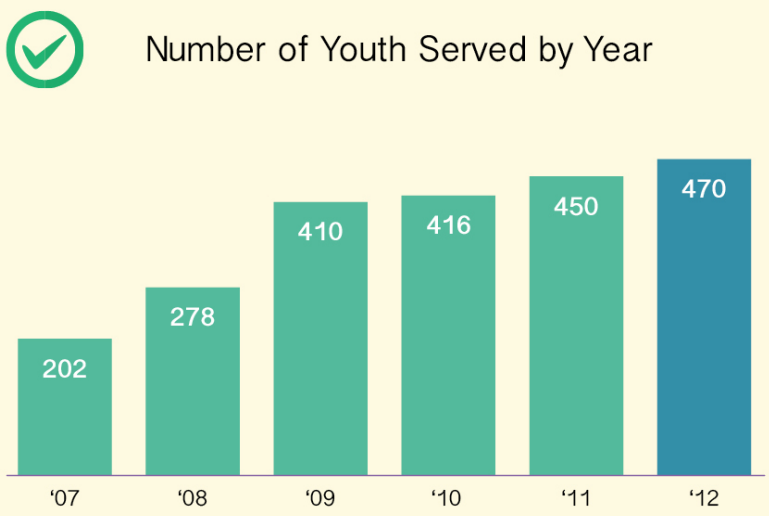
Include value annotations



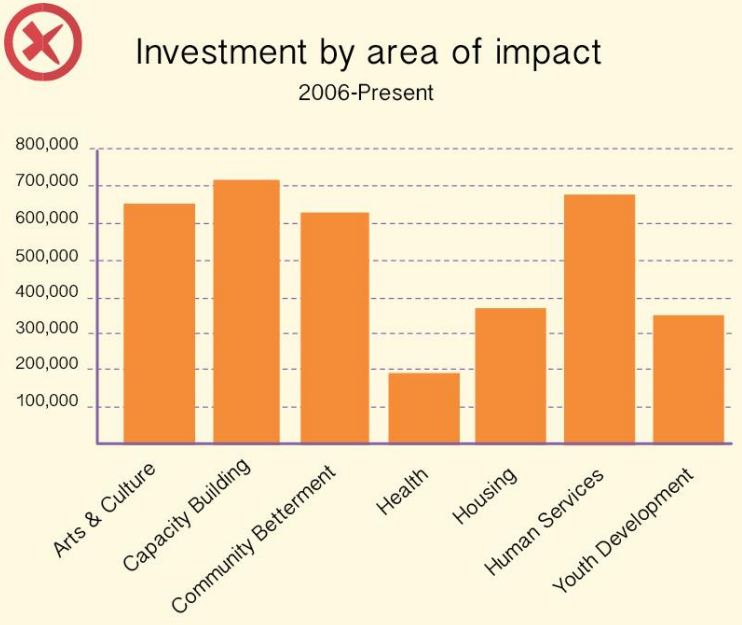
Bar Plots – Best Practices

Bar Plots – Best Practices

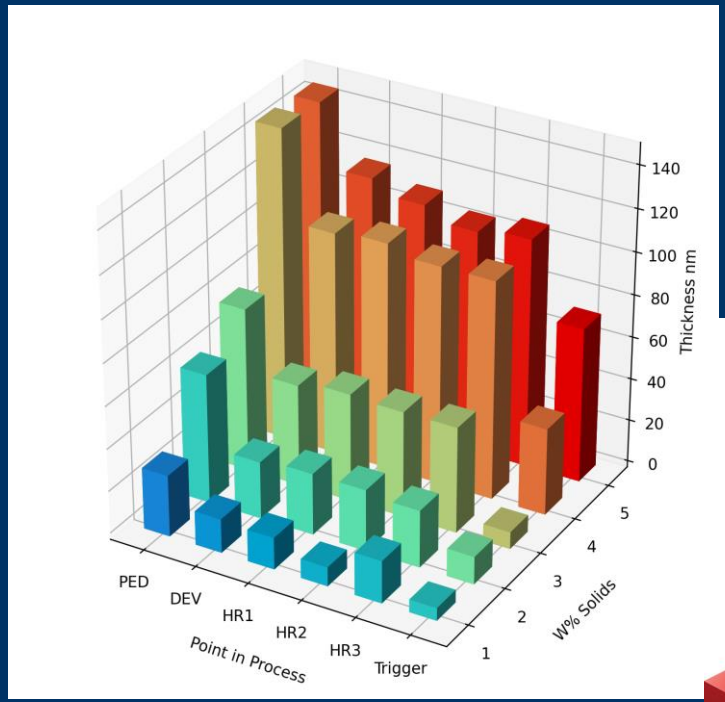


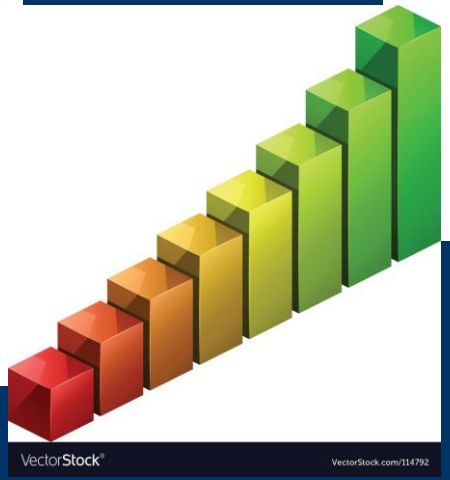


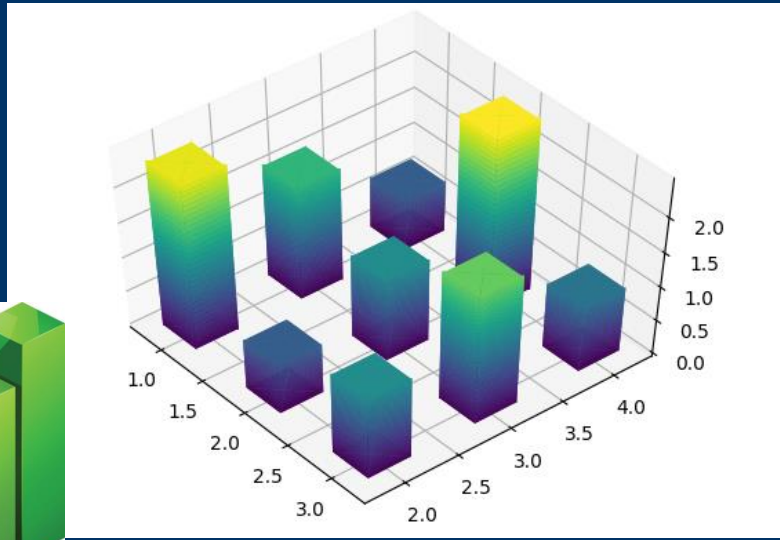
Bar Plots – Best Practices



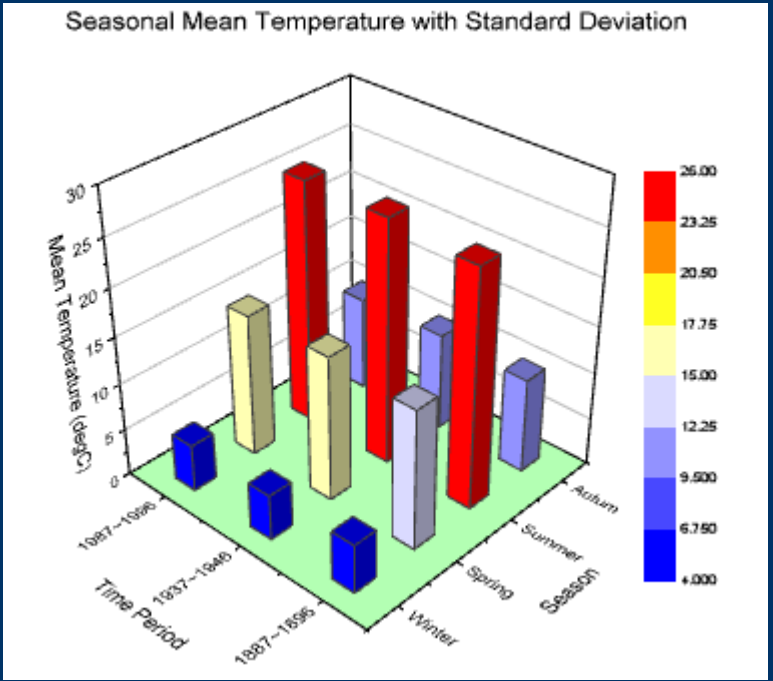
3D bar plots


140
120
100
80
60
40
20
0
Thickness nm
PED
DEV
HR1
HR2
HR3
Trigger
Point in Process
1
2
3
4
5
W% Solids

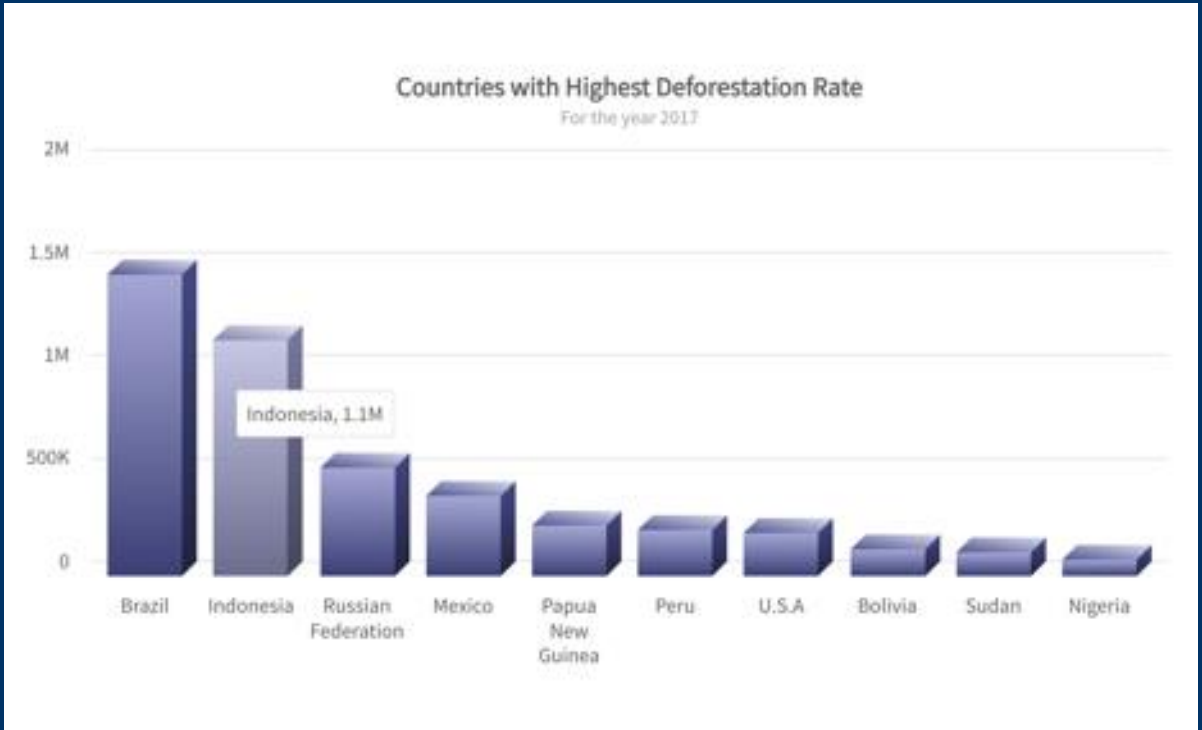

VectorStock
VectorStock.com/114792


2.0
1.5
1.0
0.5
0.0
1.0
1.5
2.0
2.5
3.0
2.0
2.5
3.0
3.5
4.0

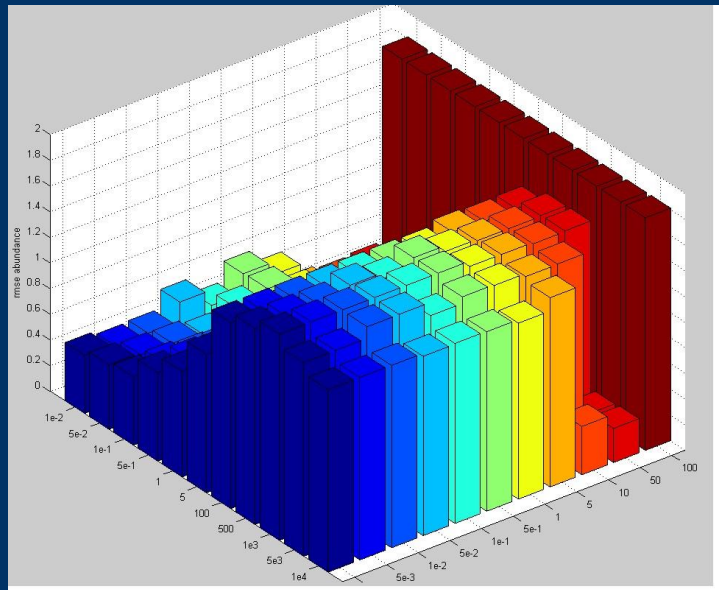
Wrong 3D Bar Plot

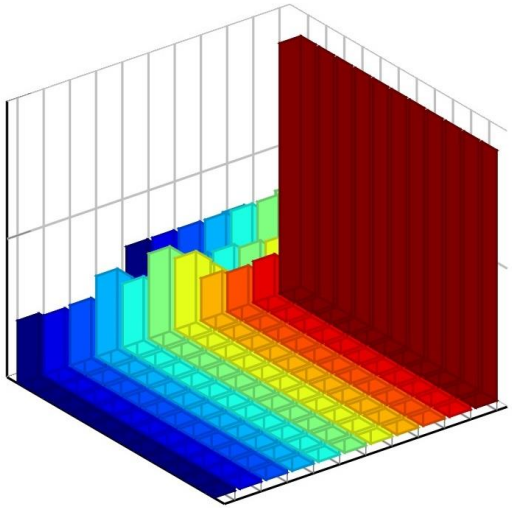


Wrong 3D Bar Plot

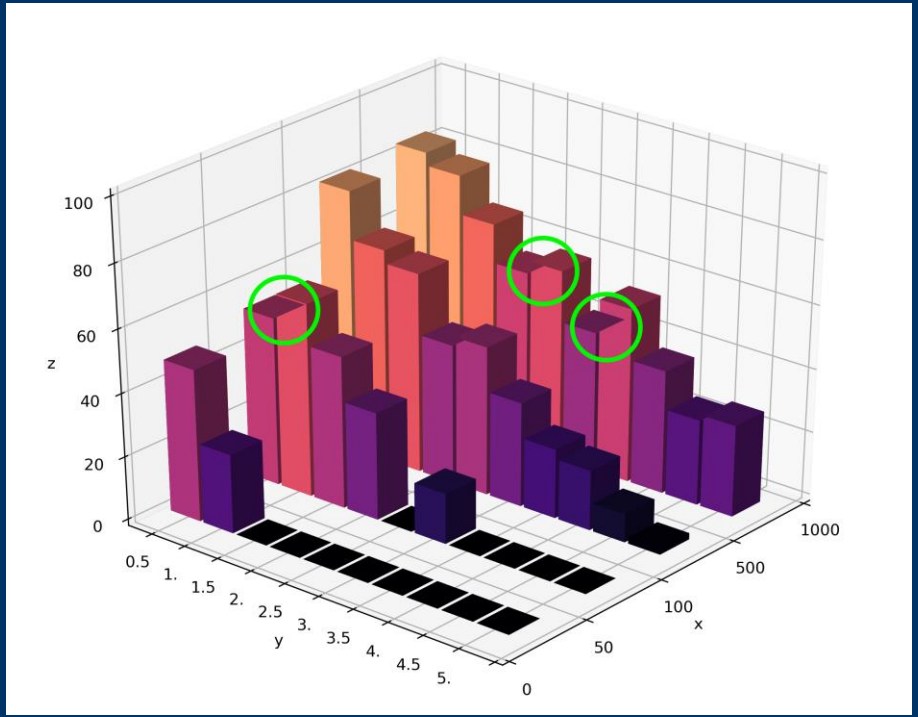


Wrong 3D Bar Plot

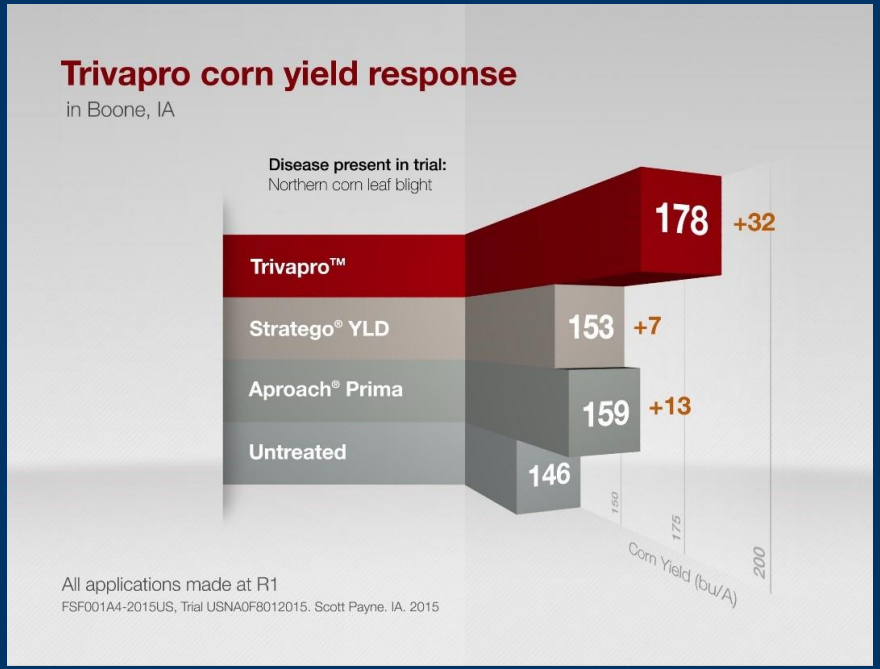




Wrong 3D Bar Plot



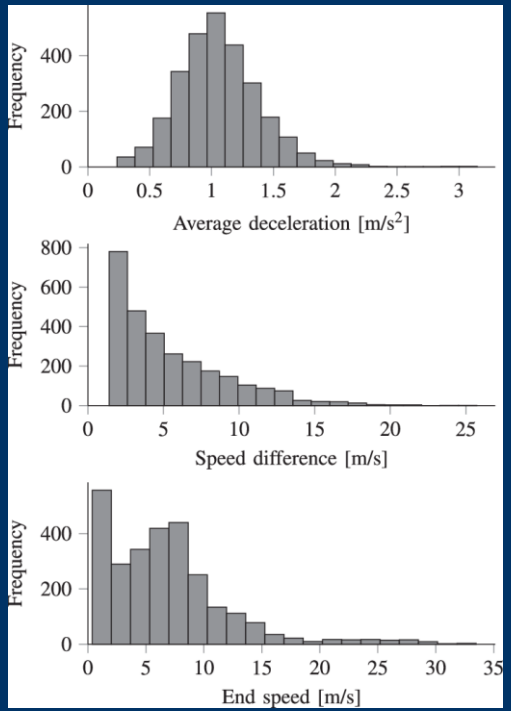
Wrong 3D Bar Plot

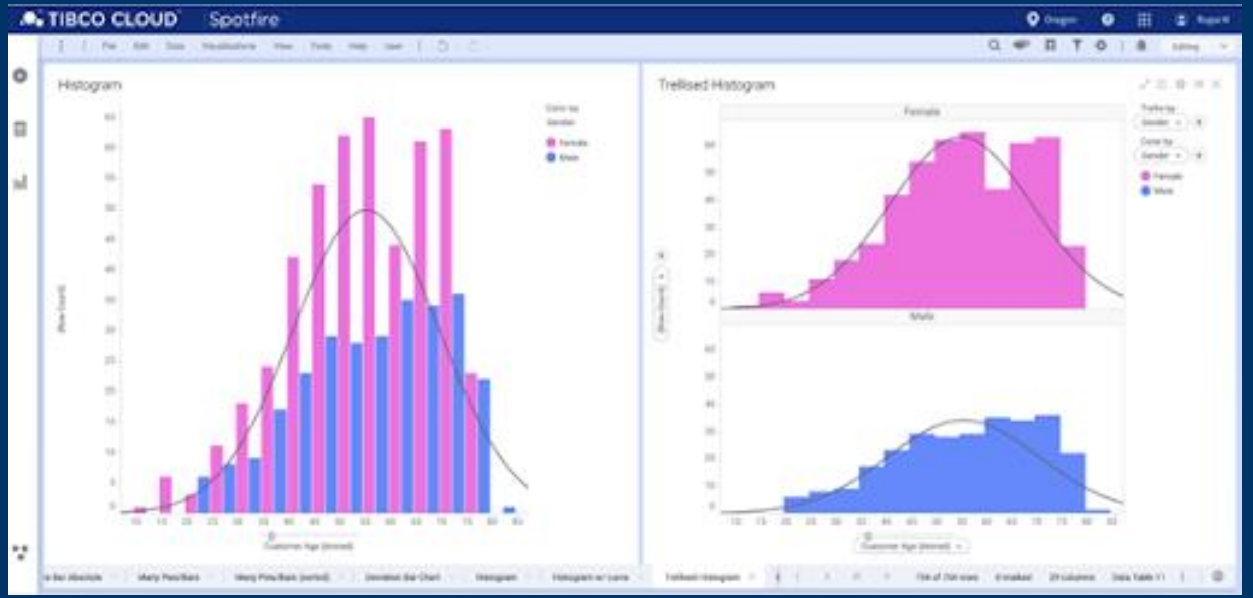


Histograms

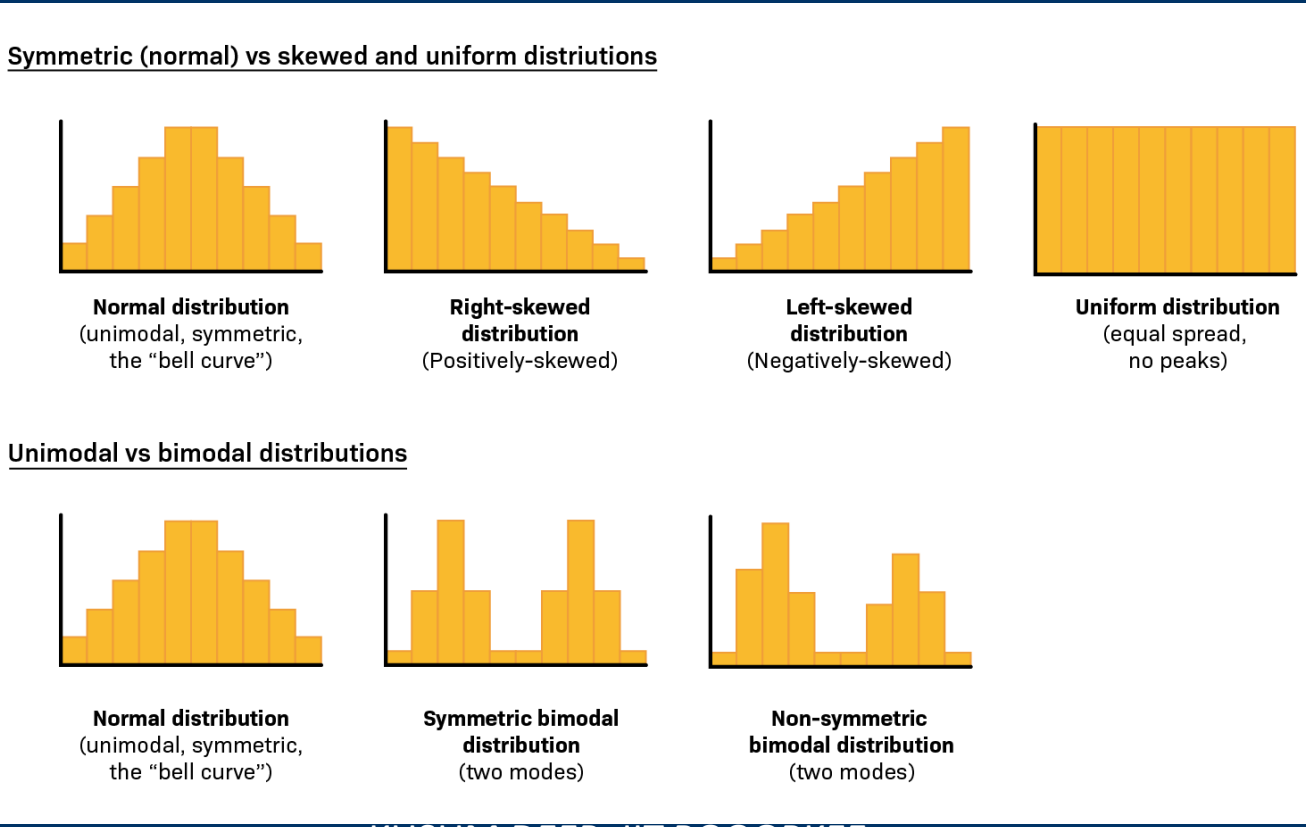
Histogram displays the distribution of numerical data. It is a graph with a series of adjacent rectangular bars, where the width of each bar represents a range of values and the height of each bar corresponds to the frequency or proportion of data within that range. Histograms are often used to identify patterns in the distribution of data, such as the presence of clusters, peaks, or gaps. They can also be used to compare the distribution of data in different groups or to track changes over time.

Histograms

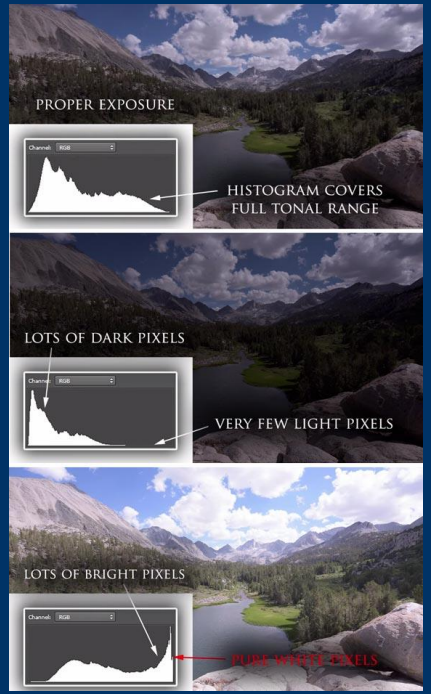

Frequency
400
200
0
0 0.5 1 1.5 2 2.5 3
Average deceleration [m/s^2]
Frequency
800
600
400
200
0
0 5 10 15 20 25
Speed difference [m/s]
Frequency
400
200
0
0 5 10 15 20 25 30 35
End speed [m/s]

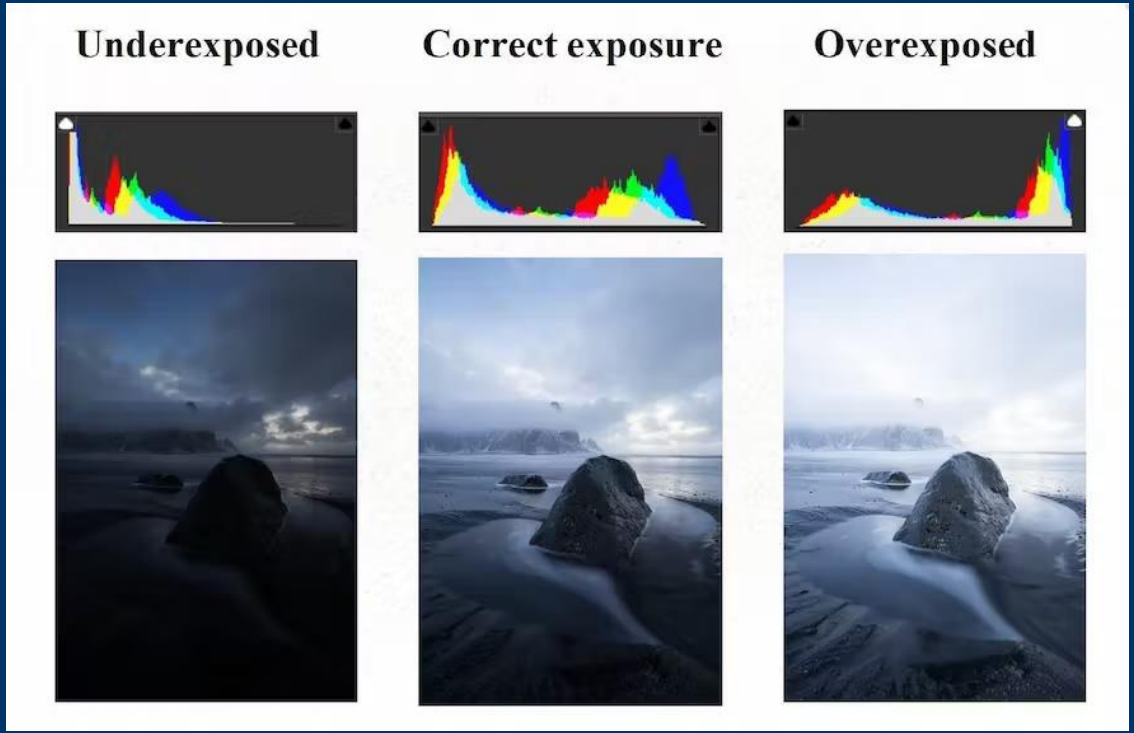

TIBCO CLOUD Spotfire
Histogram
Trellised Histogram

Histograms


Symmetric (normal) vs skewed and uniform distriutions
Normal distribution
(unimodal, symmetric,
the "bell curve")
Right-skewed
distribution
(Positively-skewed)
Left-skewed
distribution
(Negatively-skewed)
Uniform distribution
(equal spread,
no peaks)
Unimodal vs bimodal distributions
Normal distribution
(unimodal, symmetric,
the "bell curve")
Symmetric bimodal
distribution
(two modes)
Non-symmetric
bimodal distribution
(two modes)

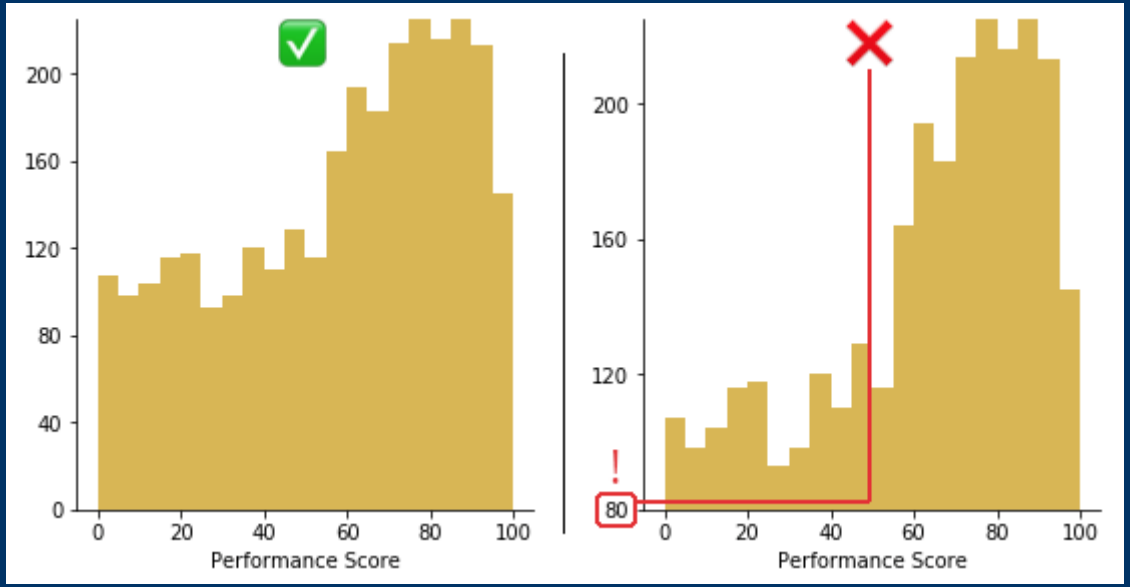
Use of histogram in photography





Histogram – Best Practices

- Use a zero-valued baseline



Bin size in histogram

Bin size in a histogram refers to the width of each bar or bin in the graph. It represents the range of values that are grouped together to create each bar in the histogram.

Choosing an appropriate bin size is important in creating an effective histogram. If the bin size is too small, the histogram may appear too "spiky" or detailed, making it difficult to identify the overall distribution pattern. If the bin size is too large, the histogram may be too smooth and may fail to capture important details in the distribution.

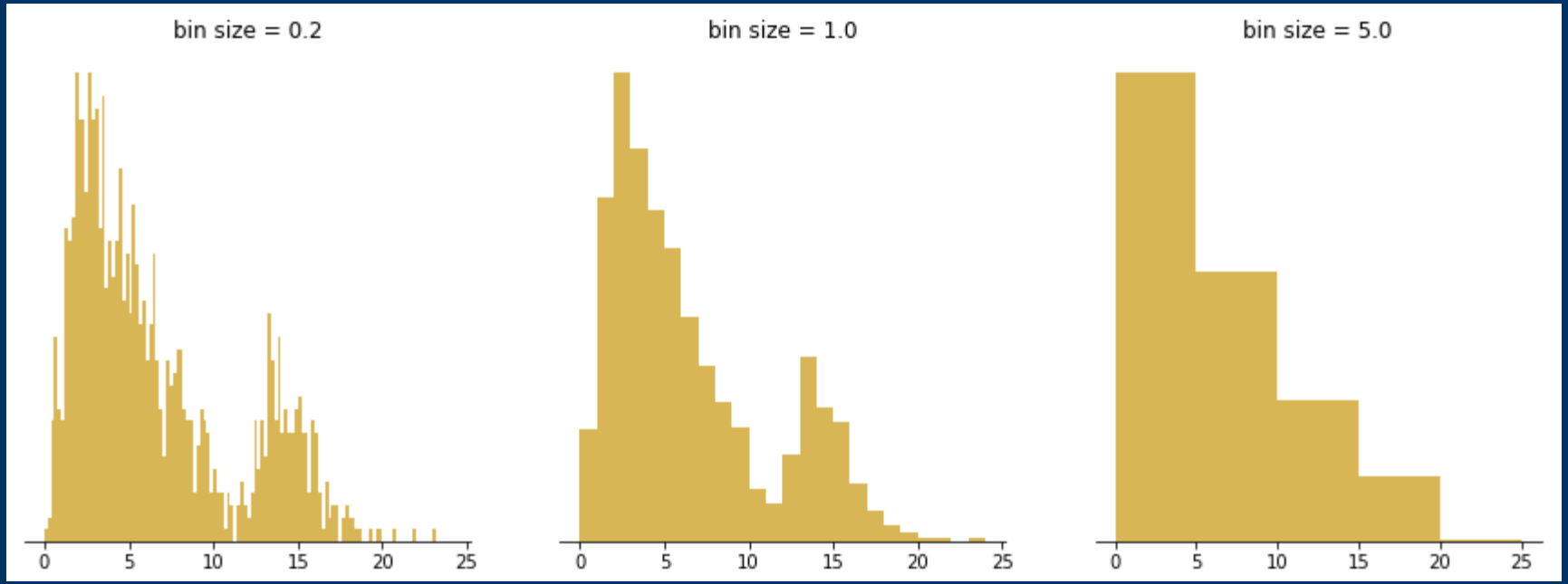
Bin size in histogram

Methods for determining the appropriate bin size for a histogram.

- Square-root rule
- Sturges' rule
- Freedman-Diaconis rule.

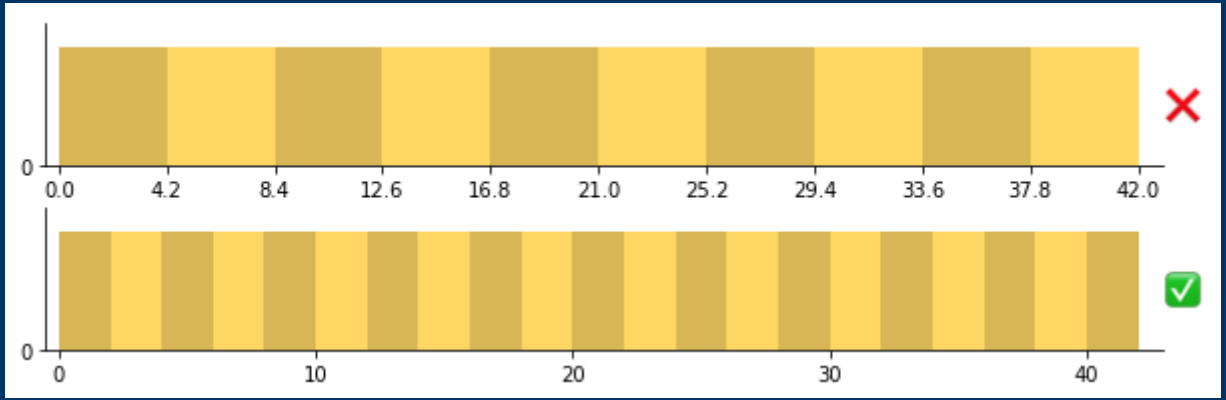
In general, the choice of bin size depends on the range of values in the data set and the objectives of the analysis. A good rule of thumb is to choose a bin size that allows for a clear and informative visualization of the data while also being easy to interpret.

Histogram – Best Practices



Histogram – Best Practices

Choose interpretable bin boundaries



Ex: Suppose we have a dataset of 1000 students' scores on a math exam. We want to visualize the distribution of scores using a histogram. The scores range from 0 to 100, and the data is stored in a list called scores.

We can create a histogram to visualize this data as follows:

```
import matplotlib.pyplot as plt
import numpy as np
scores = np.random.normal(loc=70, scale=10, size=1000)  # generate some random data
plt.hist(scores, bins=20, color='green', alpha=0.75)
plt.xlabel('Score')
plt.ylabel('Frequency')
plt.title('Distribution of Math Exam Scores')
plt.show()
```

This will produce a histogram with 20 bins, where each bin represents a range of scores. The height of each bar represents the number of students who scored within that range of scores. The x-axis shows the range of scores, and the y-axis shows the frequency (i.e., the number of students) for each range. The histogram is colored green and the opacity is set to 0.75 to make it easier to see overlapping bars.

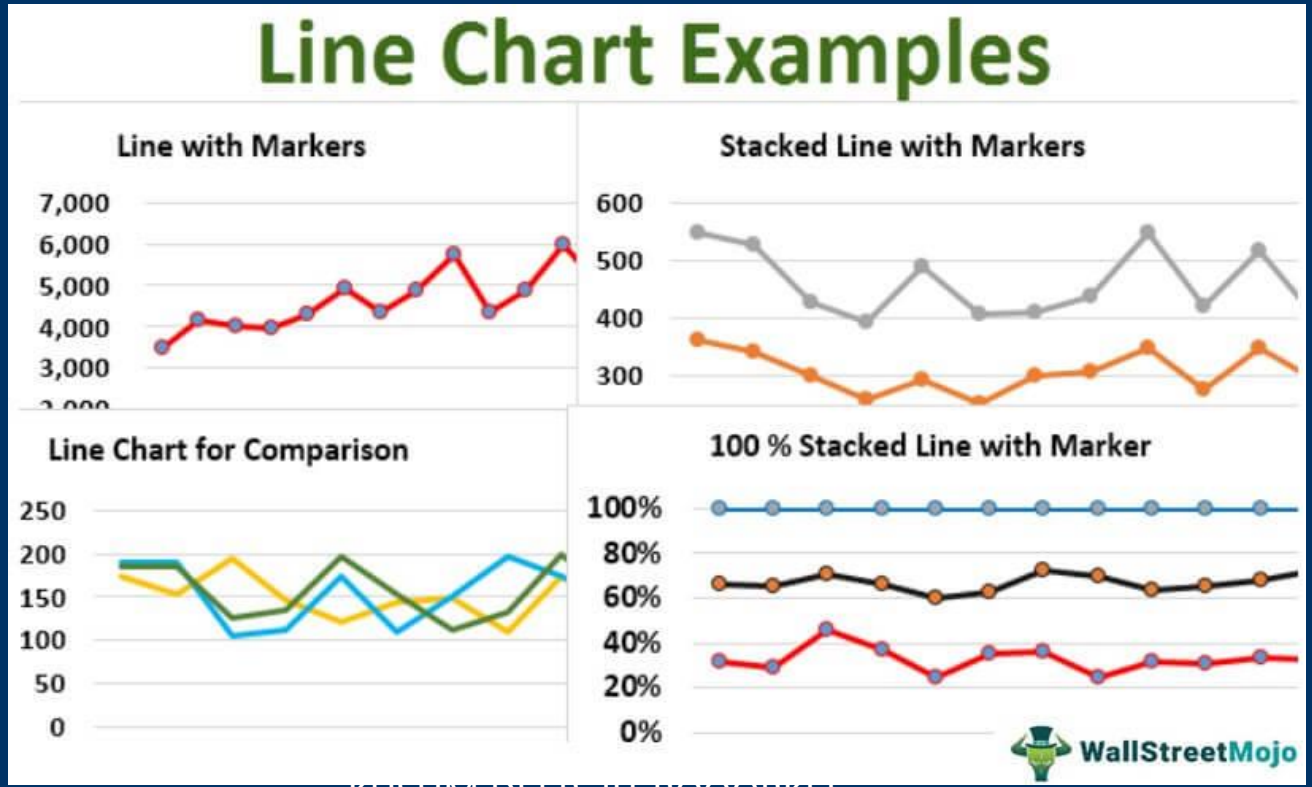
Line plots

A line plot, also known as a line graph, displays the relationship between two or more quantitative variables. The series of data points are connected by a line. Each data point represents a value for one variable, while the position of the point along the x-axis represents a value for another variable.

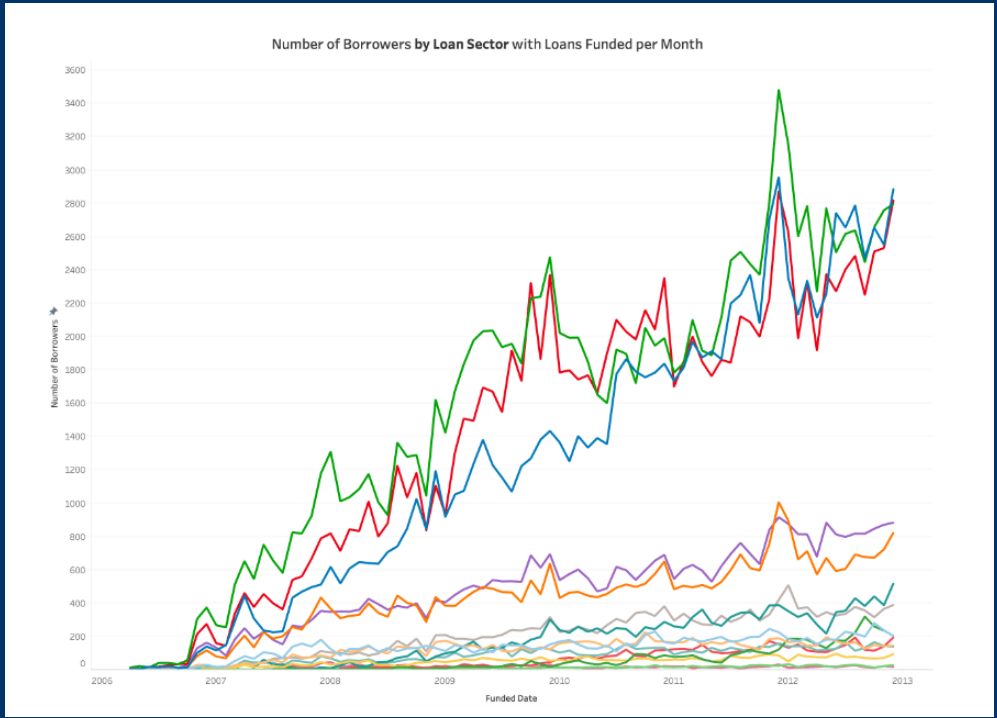
Line plots are used to show changes in data over time or to compare trends in different groups or populations. They can also be used to highlight patterns or relationships in the data, such as increasing or decreasing trends, seasonal patterns, or cyclical fluctuations.

Line plots can be customized with different colors, labels, and other visual elements to enhance their effectiveness.

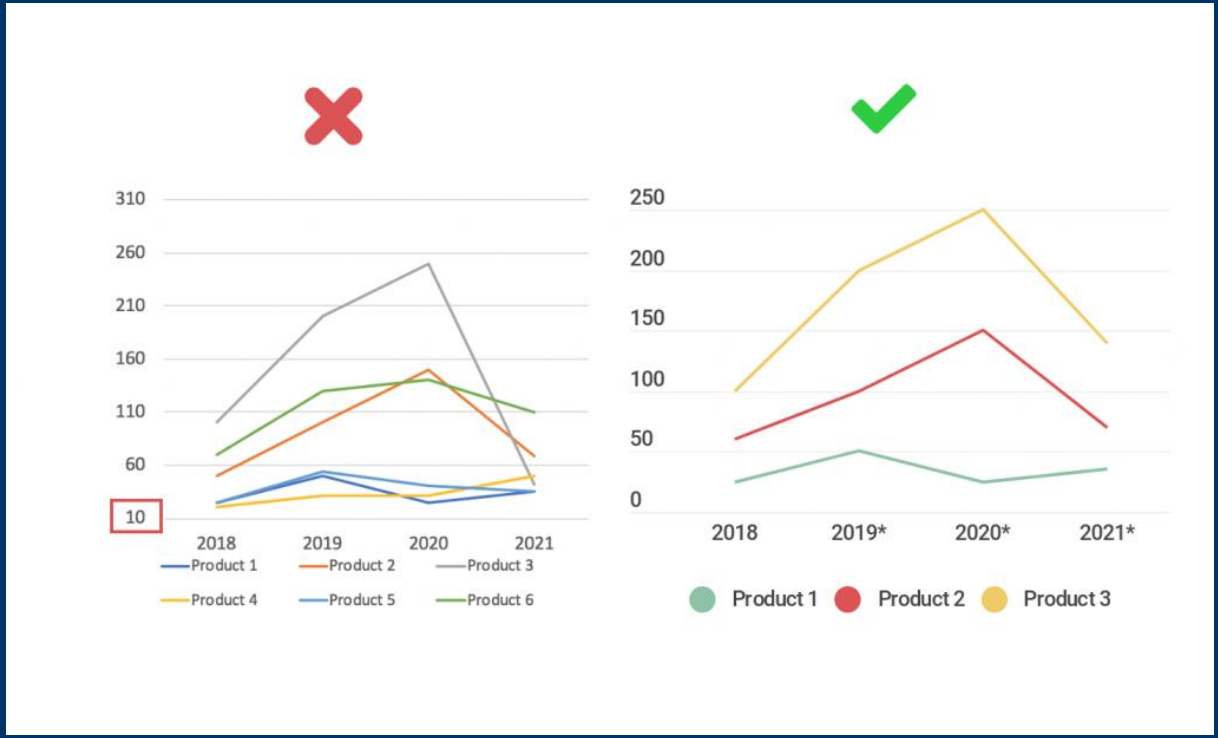
Line plots


Line Chart Examples
Line with Markers
7,000
6,000
5,000
4,000
3,000
Stacked Line with Markers
600
500
400
300
Line Chart for Comparison
250
200
150
100
50
0
100 % Stacked Line with Marker
100%
80%
60%
40%
20%
0%
WallStreetMojo

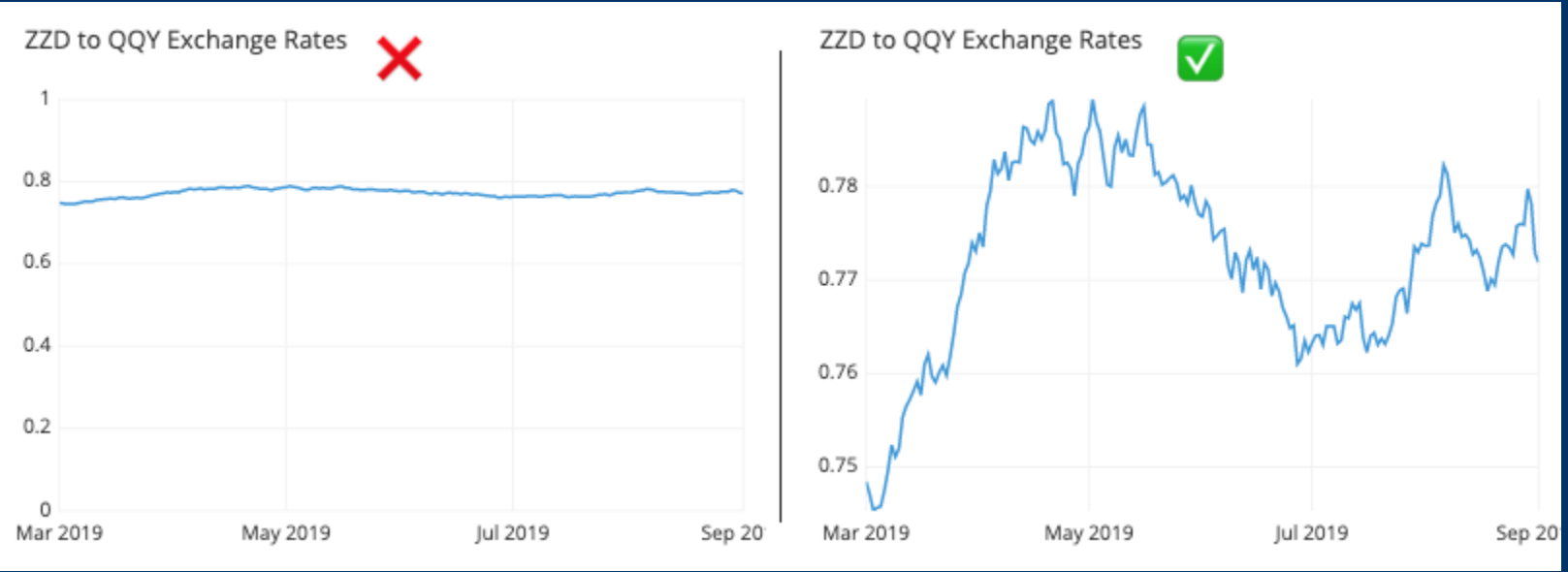
Line plots


Number of Borrowers by Loan Sector with Loans Funded per Month
Number of Borrowers
3600
3400
3200
3000
2800
2600
2400
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
2006
2007
2008
2009
2010
2011
2012
2013
Funded Date

best practices in line plots



best practices in line plots

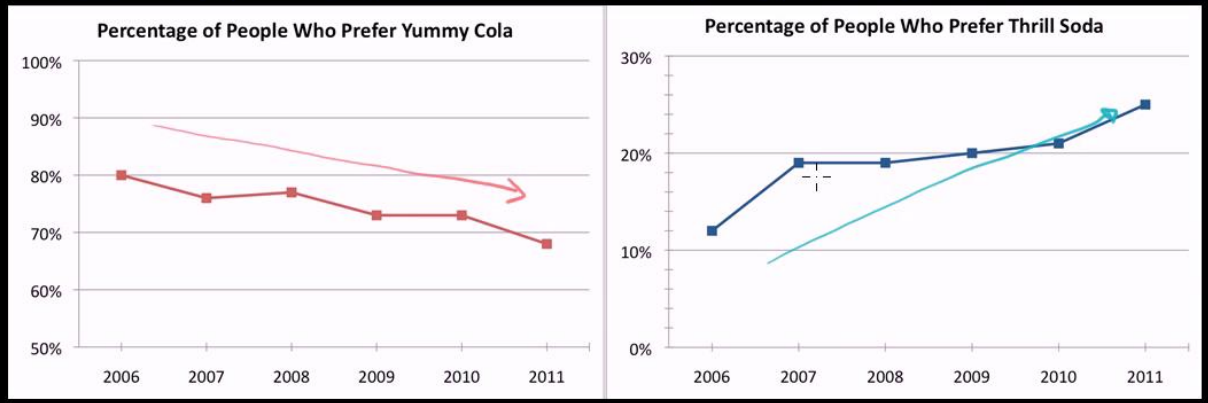


best practices in line plots

Thrill Soda hired a marketing company to help them promote their brand against Yummy Cola. The company gathered the following data about consumers' preference of soda:

Year	% of respondents who prefer Yummy Cola	% of respondents who prefer Thrill Soda	% of respondents who have no preference
2006	80%	12%	8%
2007	76%	19%	5%
2008	77%	19%	4%
2009	73%	20%	7%
2010	73%	21%	6%
2011	68%	25%	7%

The advertising company created the following two graphs to promote Thrill Soda:

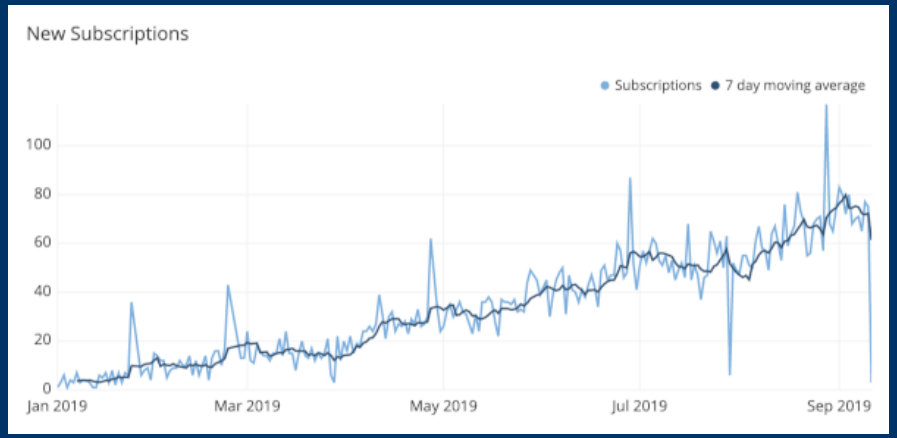


khanacademy.org

Line plot – Best Practices

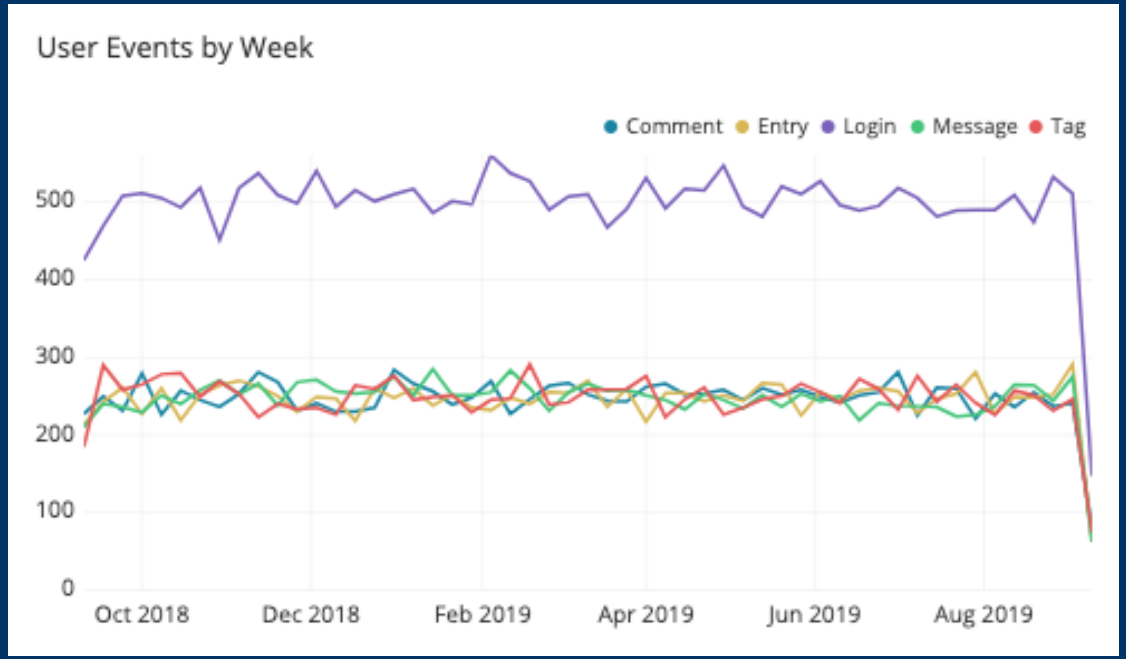
Appropriate measurement interval

If a measurement interval is too broad, it may mean that it takes too long to see where the data trend is leading and if the measurement interval is too small, it may only reveal noise rather than signal.



Line plot – Best Practices

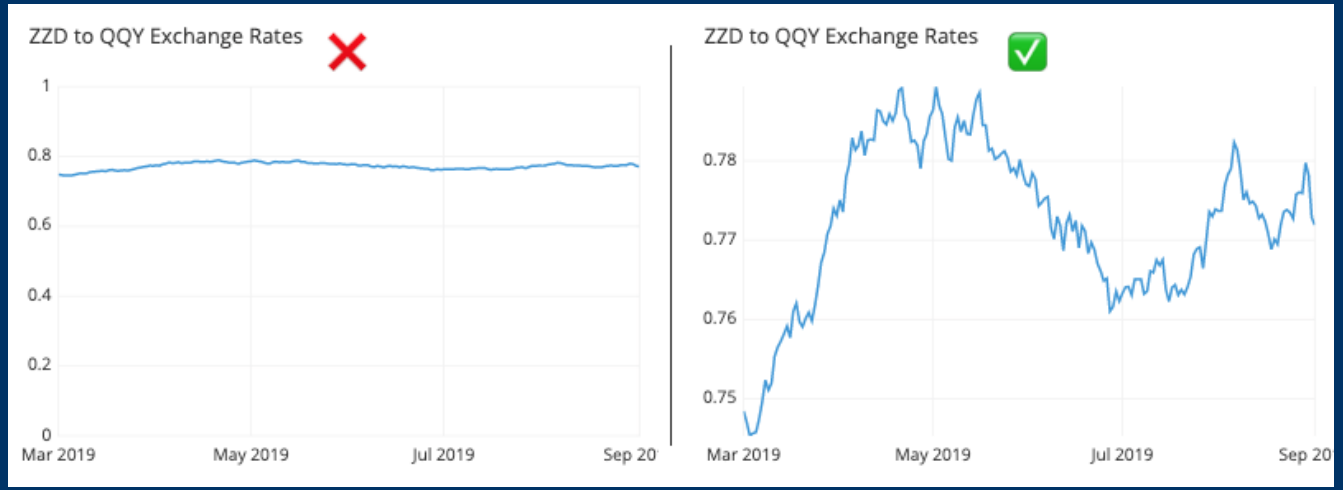
Don’t plot too many lines



Line plot – Best Practices

Zero-value baseline may not be used

Despite the zero baseline for the vertical axis being a requirement for bar charts and histograms, you do not need to include a zero baseline for a line chart.



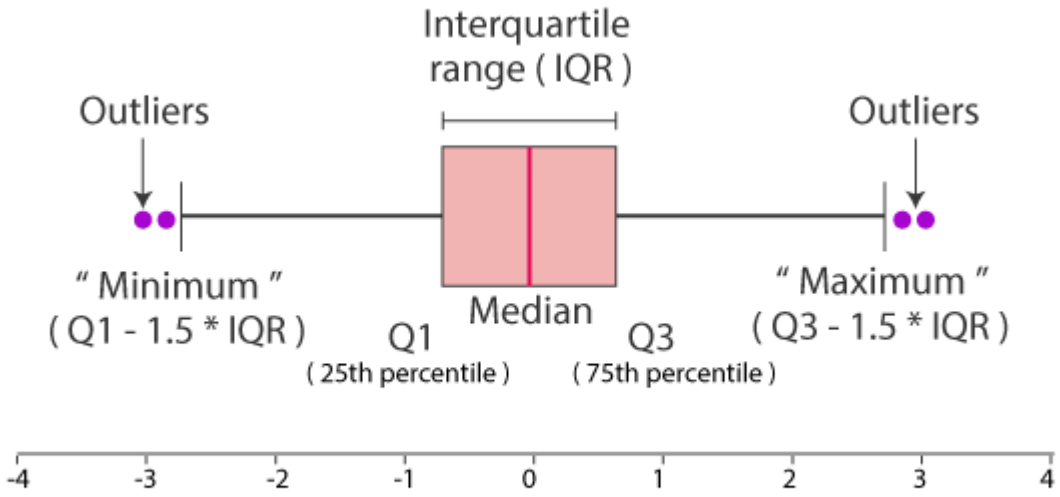
Box plots

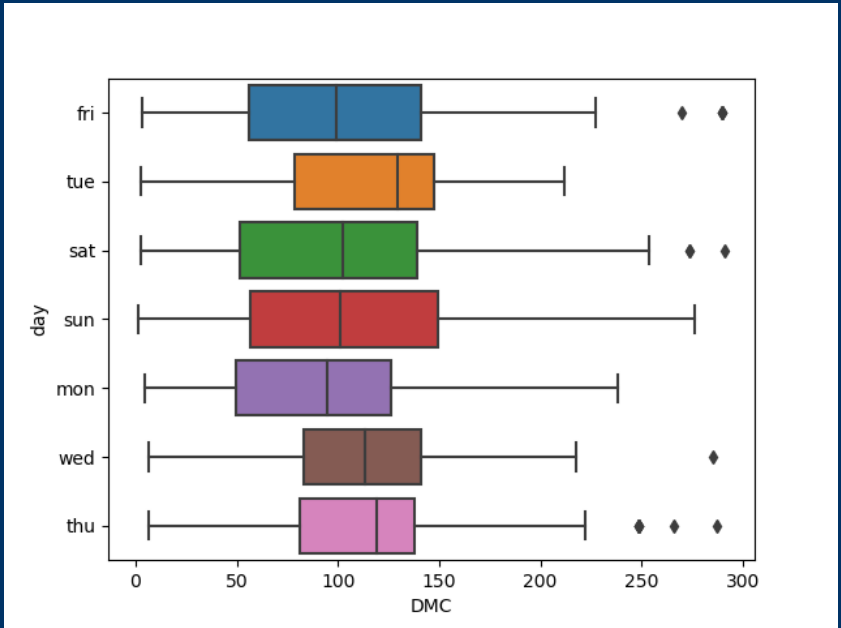
Box plots (box-and-whisker plots) are a graphical representation of a set of data through five key summary statistics: minimum value, first quartile (Q1), median, third quartile (Q3), and maximum value.

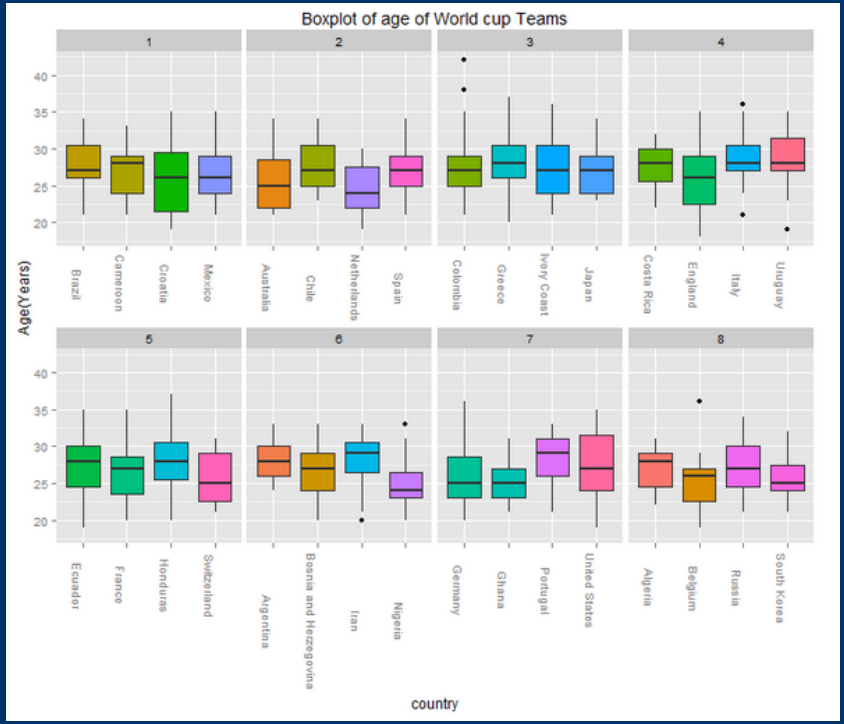
The box in the middle of the plot represents the interquartile range (IQR), which is the range of values between Q1 and Q3.

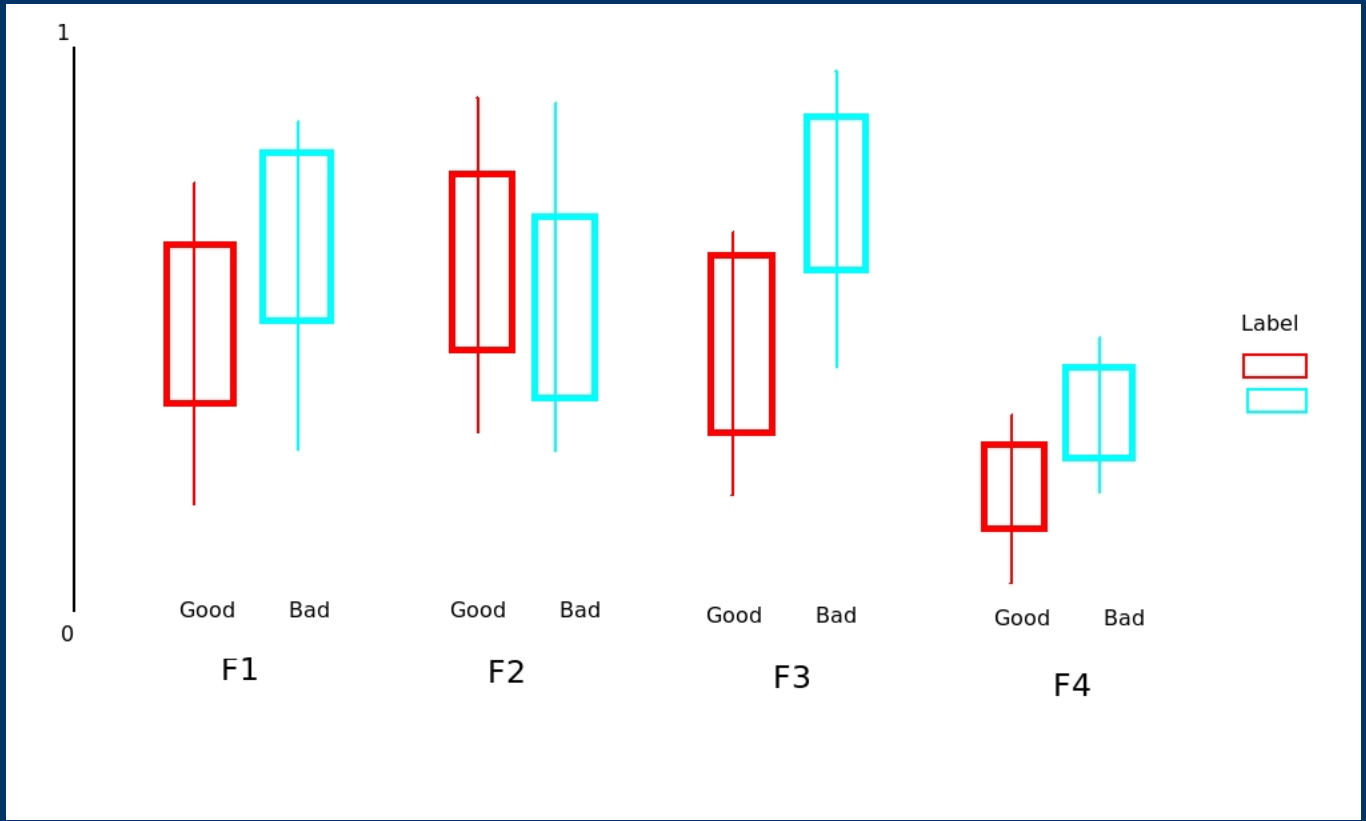
The whiskers, which extend from the box, represent the range of data outside the IQR. It can also show outliers, which are data points that fall outside of the whiskers. Box plots are useful for comparing the distribution of data between multiple groups or for visualizing the distribution of a single variable.

Box Plots


Interquartile
range (IQR)
Outliers
Outliers
" Minimum "
(Q1 - 1.5 * IQR)
Q1
(25th percentile)
Median
Q3
(75th percentile)
" Maximum "
(Q3 - 1.5 * IQR)
-4
-3
-2
-1
0
1
2
3
4
Different parts of boxplot
© Byjus.com


fri
tue
sat
sun
mon
wed
thu
day
0
50
100
150
200
250
300
DMC


Boxplot of age of World cup Teams
1
2
3
4
5
6
7
8
Age(Years)
Brazil
Cameroon
Croatia
Mexico
Australia
Chile
Netherlands
Spain
Colombia
Greece
Ivory Coast
Japan
Costa Rica
England
Italy
Uruguay
Ecuador
France
Honduras
Switzerland
Argentina
Bosnia and Herzegovina
Iran
Nigeria
Germany
Ghana
Portugal
United States
Algeria
Belgium
Russia
South Korea
country

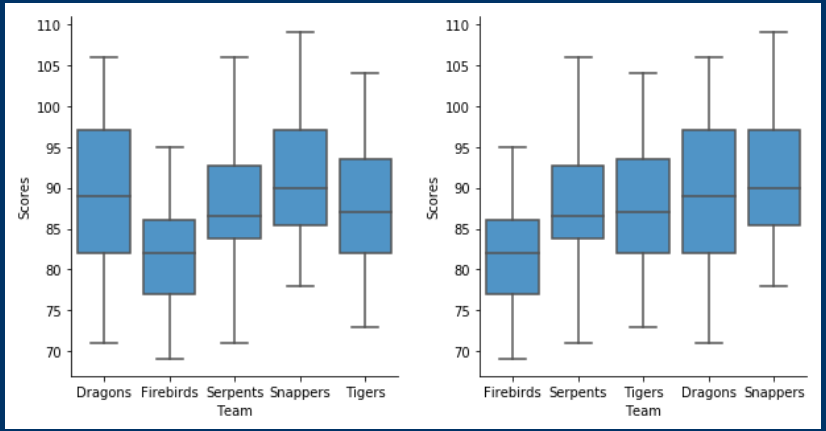

1
0
Good
Bad
F1
Good
Bad
F2
Good
Bad
F3
Good
Bad
F4
Label

Box plots – Best Practices

Consider the order of groups

If the groups plotted in a box plot do not have an inherent order, then arrange them in an order that highlights patterns and insights.

Popular ordering for groups is to sort them by median value.



Suppose we have a dataset of 500 students' scores on a math exam. We want to visualize the distribution of scores using a box plot. The scores range from 0 to 100, and the data is stored in a list called scores. We can create a box plot to visualize this data as follows:

```
import matplotlib.pyplot as plt
import numpy as np
scores = np.random.normal(loc=70, scale=10, size=500)  # generate some random data
plt.boxplot(scores)
plt.xlabel('Math Exam Scores')
plt.title('Distribution of Math Exam Scores')
plt.show()
```

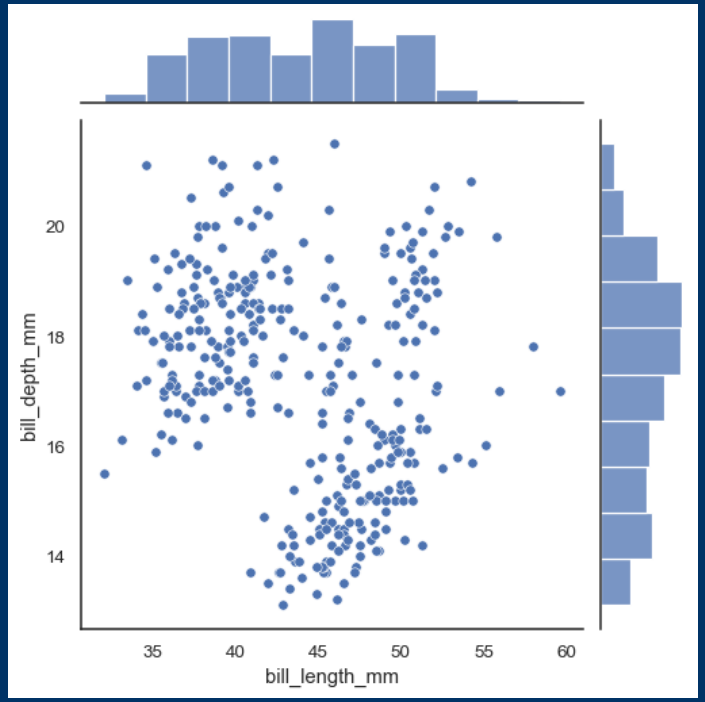
This will produce a box plot with a single box, where the box represents the interquartile range (IQR) of the data. The horizontal line inside the box represents the median (i.e., the 50th percentile) of the data. The whiskers extend from the box to show the range of the data, excluding any outliers. Outliers are plotted as individual points outside of the whiskers. The x-axis is labeled "Math Exam Scores" and the title of the plot is "Distribution of Math Exam Scores."

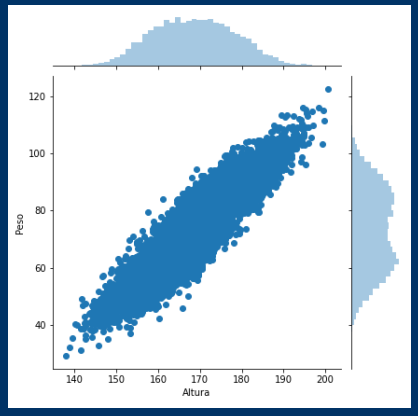
Joint plots

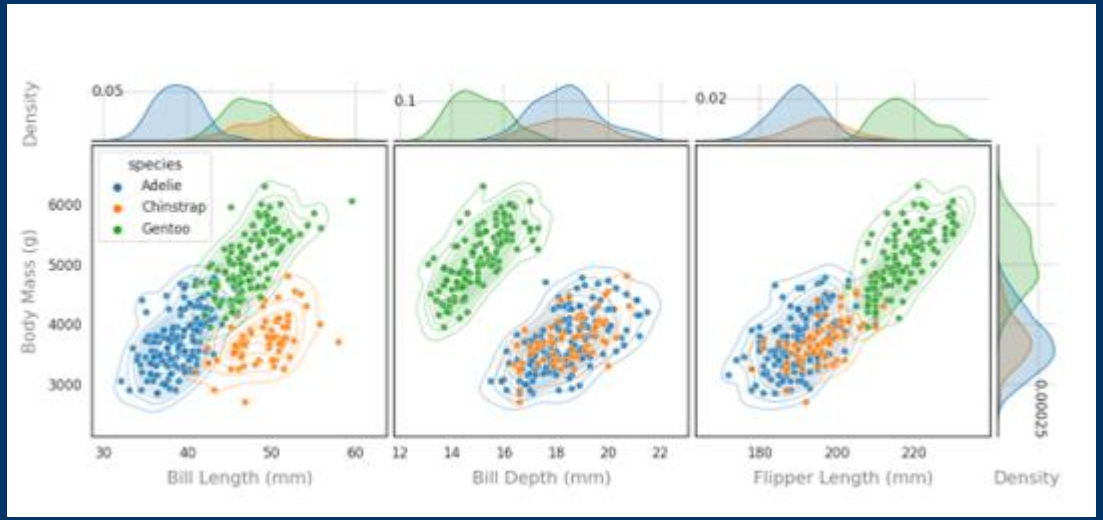
It combine scatter plots and histograms into a single plot.

The two variables of interest are plotted against each other in a scatter plot, with each data point represented by a dot. Along the x and y axes of the plot, histograms are included to show the distribution of each variable separately. The plot can also include additional information, such as regression lines or density plots. There are three plots.

- One plot displays a bivariate graph which shows how the dependent variable(Y) varies with the independent variable(X)
- Second plot is placed horizontally at the top of the bivariate graph and it shows the distribution of the independent variable(X).
- Third plot is placed on the right margin of the bivariate graph with the orientation set to vertical and it shows the distribution of the dependent variable(Y)


bill_depth_mm
bill_length_mm
20
18
16
14
35
40
45
50
55
60


Peso
Altura
120
100
80
60
40
140
150
160
170
180
190
200


Density
0.05
0.1
0.02
species
Adelie
Chinstrap
Gentoo
Body Mass (g)
6000
5000
4000
3000
30
40
50
60
12
14
16
18
20
22
180
200
220
0.00025
Bill Length (mm)
Bill Depth (mm)
Flipper Length (mm)
Density

Suppose we have a dataset of 1000 students' scores on a math exam and their corresponding scores on a science exam. We want to visualize the relationship between the two sets of scores using a joint plot. The scores range from 0 to 100, and the data is stored in two lists called math_scores and science_scores. We can create a joint plot to visualize this data as follows:

```
import seaborn as sns
import numpy as np
math_scores = np.random.normal(loc=70, scale=10, size=1000)  # generate some random data
science_scores = np.random.normal(loc=70, scale=10, size=1000)
sns.jointplot(x=math_scores, y=science_scores, kind='scatter’)
```

This will produce a joint plot with two histograms and a scatter plot. The histograms show the distribution of math scores and science scores separately, while the scatter plot shows the relationship between the two sets of scores. The x-axis shows math scores, the y-axis shows science scores, and the title of the plot is "Joint Plot of Math Scores and Science Scores". The kind parameter is set to "scatter" to produce a scatter plot. Other options for kind include "hex" for a hexbin plot, and "kde" for a 2D kernel density estimate plot.

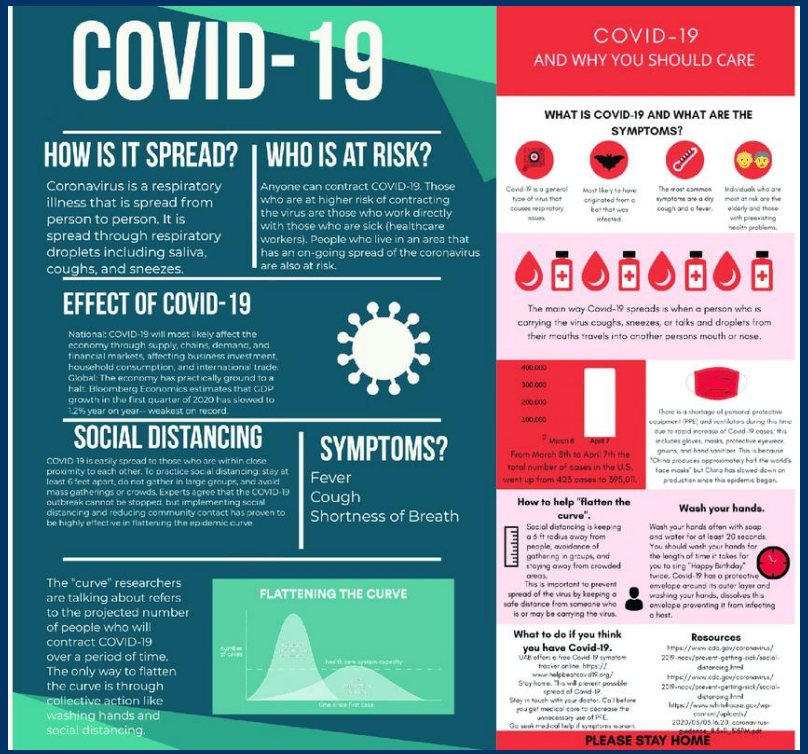
infographics

They are visual representations of data and information that aim to communicate complex ideas in a clear and concise way. Infographics can take many different forms, including charts, diagrams, maps, timelines, and more.

Their goal is to present data or information in a way that is easy to understand and visually appealing. Infographics typically use color, typography, and other design elements to highlight key points and make the information more engaging.

Used in journalism, marketing, education, and research.

Infographics




6 Steps To Help You Save Money
Keep track of your expenses.
Make saving a priority in your budget.
Look for ways to save money.
Set financial objectives.
Establish your financial priorities.
Observe your savings grow


What is an Infographic?
VENNGAGE


HOW TO PROPERLY WASH YOUR HAND
1
Wash your hands with water and soap
2
Lather well beyond your wrists
3
Wash all surfaces thoroughly including wrists, palms and under fingers
4
Rub hands together for at least 30 seconds
5
Rinse thoroughly and dry completely

Ex: Use Infographics to design:

1. 12 th International Conference on Soft Computing for Problem Solving to be held at IIT Roorkee during August 11-13, 2023. http://www.socpros2023.iitr.ac.in/

2.

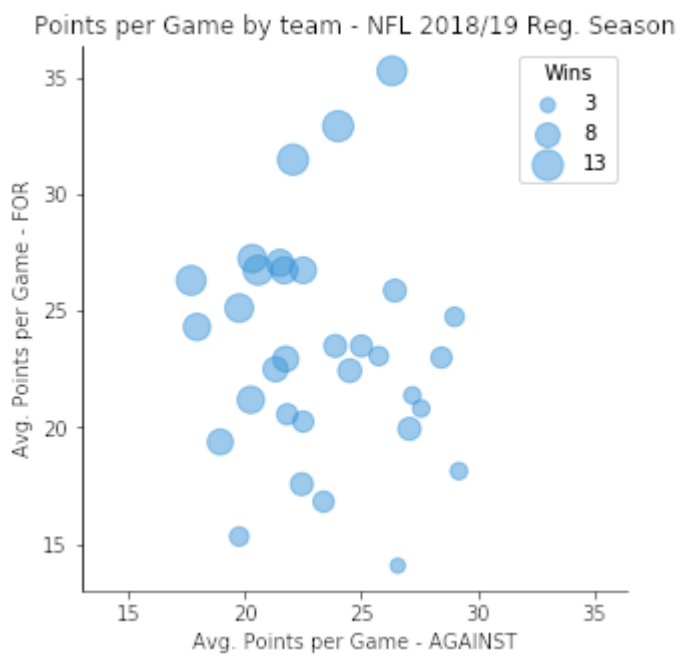
3.

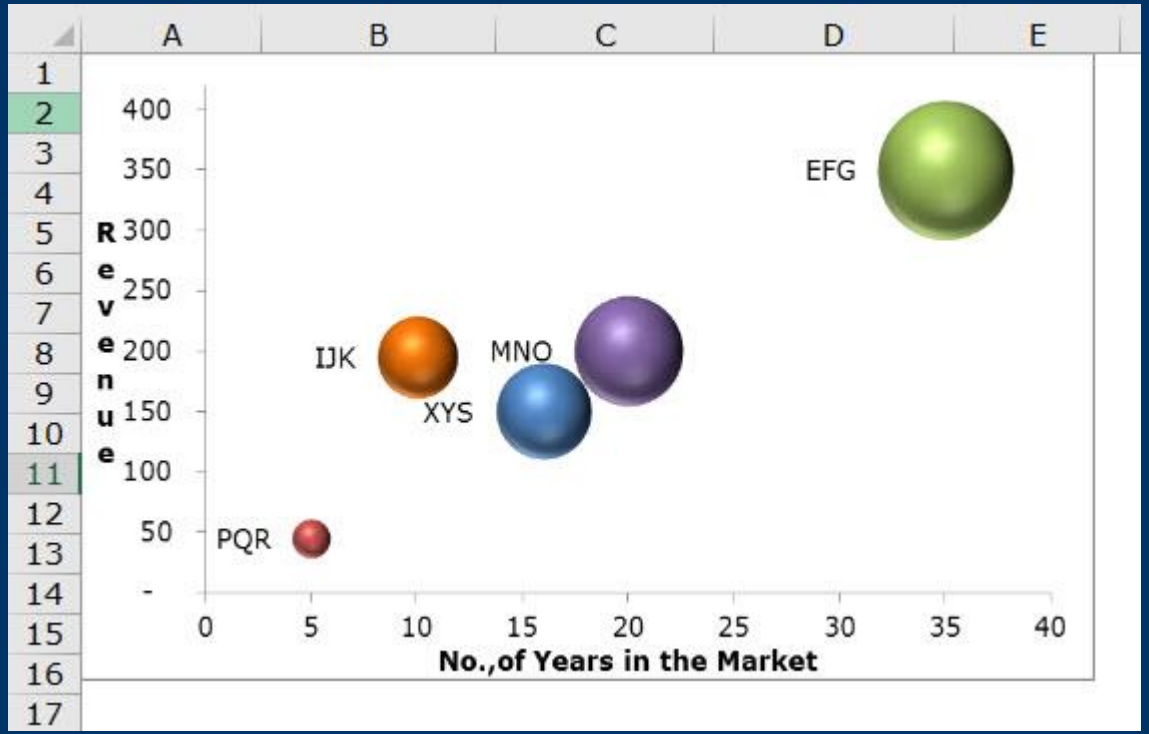
4.

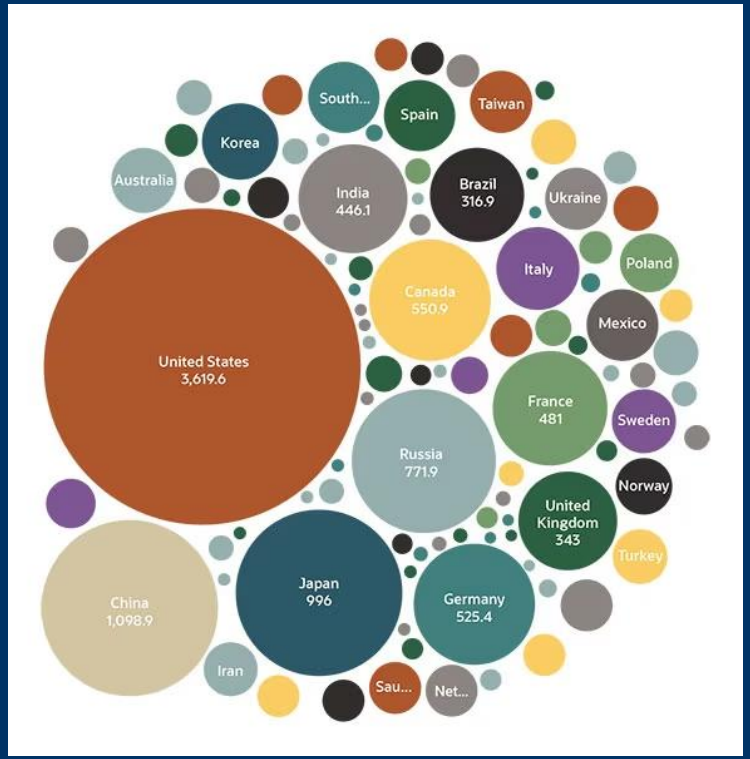
Bubble charts

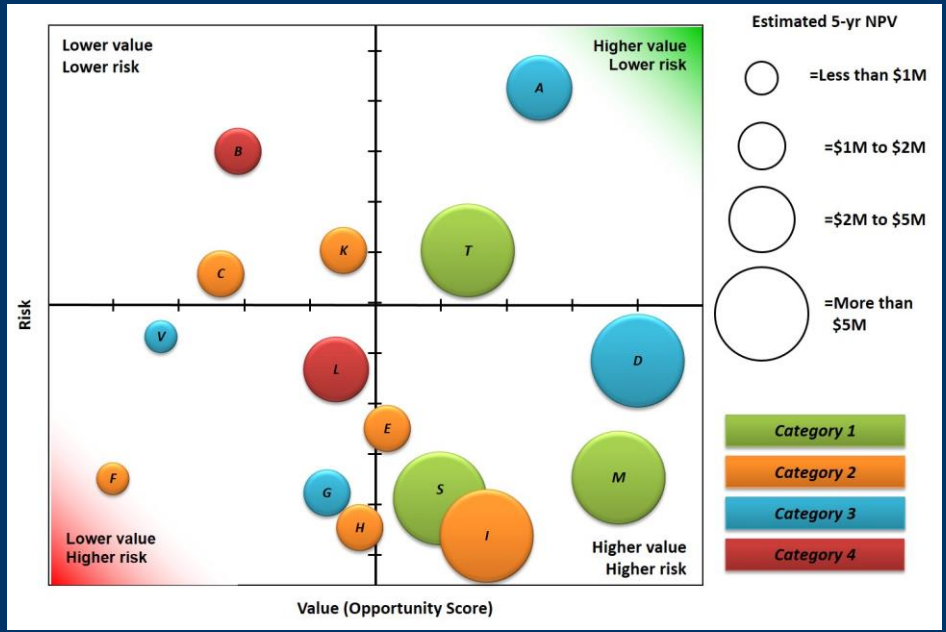
It display data points as bubbles on a two-dimensional plot. Each bubble represents a single data point and is positioned according to its x and y values. Additionally, the size of the bubble is proportional to a third variable, allowing for the visualization of three dimensions of data in a single chart.

Bubble charts are useful for comparing and visualizing data across multiple variables, as well as for identifying patterns and trends in the data.

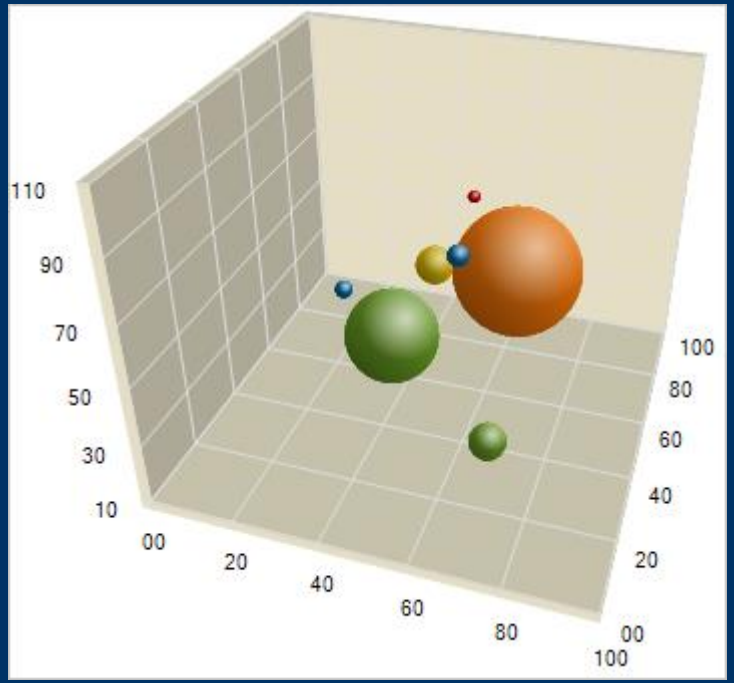

Points per Game by team - NFL 2018/19 Reg. Season
Wins
3
8
13
Avg. Points per Game - FOR
Avg. Points per Game - AGAINST
35
30
25
20
15
15
20
25
30
35

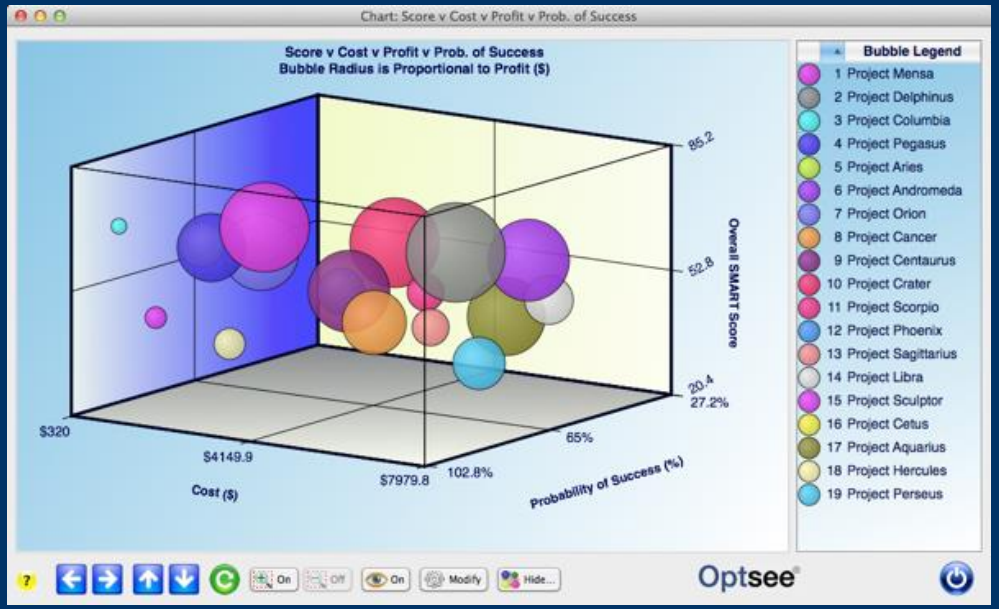

A
B
C
D
E
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Revenue
400
350
300
250
200
150
100
50
-
EFG
IJK
MNO
XYS
PQR
0
5
10
15
20
25
30
35
40
No.,of Years in the Market


South...
Spain
Taiwan
Korea
Australia
India
446.1
Brazil
316.9
Ukraine
Italy
Poland
Canada
550.9
Mexico
United States
3,619.6
France
481
Sweden
Russia
771.9
Norway
United
Kingdom
343
Turkey
China
1,098.9
Japan
996
Germany
525.4
Iran
Sau...
Net...


Lower value
Lower risk
Higher value
Lower risk
Lower value
Higher risk
Higher value
Higher risk
Risk
Value (Opportunity Score)
A
B
C
K
T
V
L
D
E
F
G
H
S
I
M
Estimated 5-yr NPV
=Less than $1M
=$1M to $2M
=$2M to $5M
=More than $5M
Category 1
Category 2
Category 3
Category 4

3D Bubble chart





Suppose we have a dataset of 50 companies and their corresponding revenue, profit, and market capitalization. We want to visualize the relationship between these three variables using a 3D bubble chart. The data is stored in a pandas DataFrame called companies.

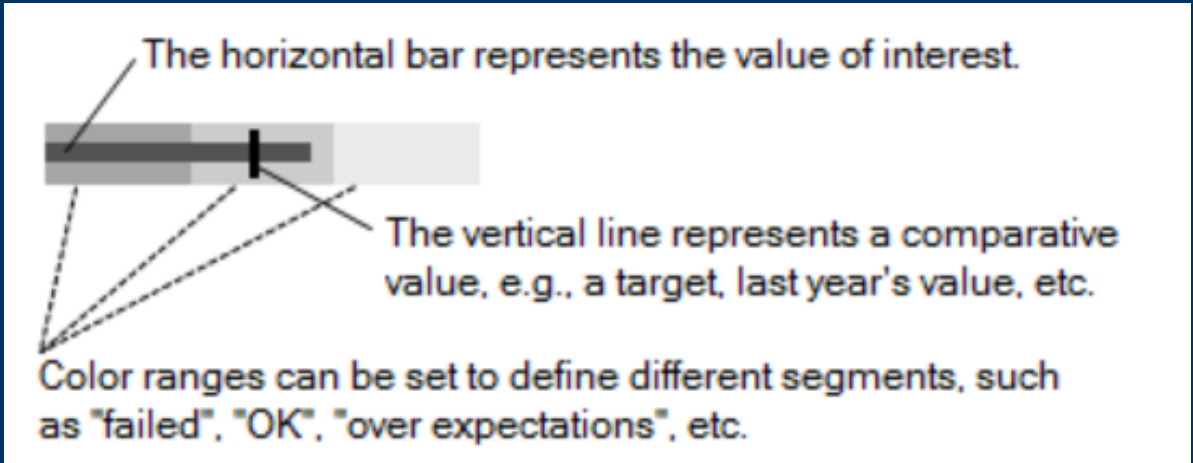
We can create a 3D bubble chart to visualize this data as follows:

```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
from mpl_toolkits.mplot3d import Axes3D
# create a DataFrame with random data
np.random.seed(42)
companies = pd.DataFrame({
    'revenue': np.random.randint(low=1000, high=10000, size=50),
    'profit': np.random.randint(low=50, high=500, size=50),
    'market_cap': np.random.randint(low=5000, high=50000, size=50)
})
# create a 3D plot
fig = plt.figure(figsize=(10, 8))
ax = fig.add_subplot(111, projection='3d')
ax.scatter(companies['revenue'], companies['profit'], companies['market_cap'], s=companies['market_cap']/1000, alpha=0.8)
# set labels and title
ax.set_xlabel('Revenue')
ax.set_ylabel('Profit')
ax.set_zlabel('Market Cap')
ax.set_title('3D Bubble Chart of Company Metrics')
plt.show()
```

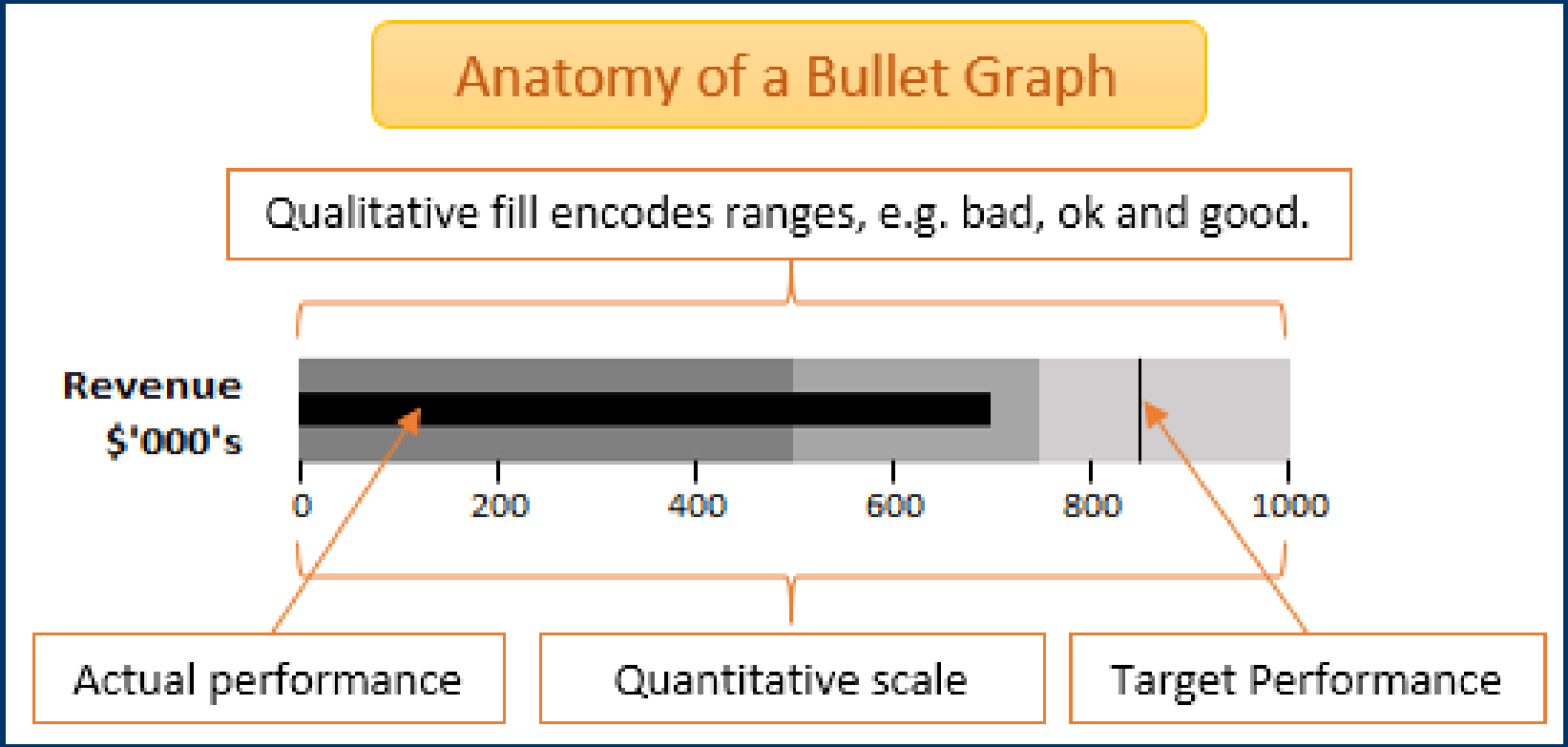
This will produce a 3D bubble chart with bubbles of varying size representing market capitalization. The x-axis shows revenue, the y-axis shows profit, and the z-axis shows market capitalization. The size of each bubble is proportional to the company's market capitalization, and the opacity is set to 0.8 to make the chart easier to read. The labels on the axes and the title of the chart are set using ax.set_xlabel(), ax.set_ylabel(), `ax.set

Bullet graphs

A bullet graph is a variation of a bar graph developed to replace dashboard gauges and meters. It is useful for comparing the performance of a primary measure to one or more other measures.



Bullet graph



Ex: Suppose we want to visualize the performance of a manufacturing plant on three different quality metrics: defect rate, on-time delivery, and yield. We can use a 3D bullet graph to display this information.

First, we need to define the target, actual, and comparative values for each metric:

Defect rate: Target = 2%, Actual = 1.5%, Comparative = 3%

On-time delivery: Target = 95%, Actual = 90%, Comparative = 85%

Yield: Target = 98%, Actual = 97%, Comparative = 95%

We can then plot these values on a 3D chart, where each metric is represented by a different axis:

The x-axis represents defect rate

The y-axis represents on-time delivery

The z-axis represents yield

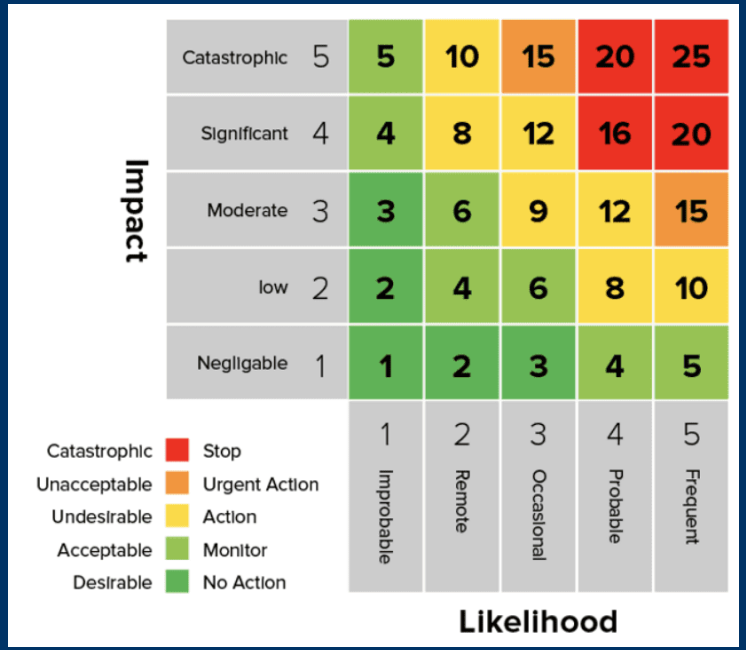
Plot each value as a point in 3D space, and connect the actual value to the target and comparative values using a series of lines or bars.

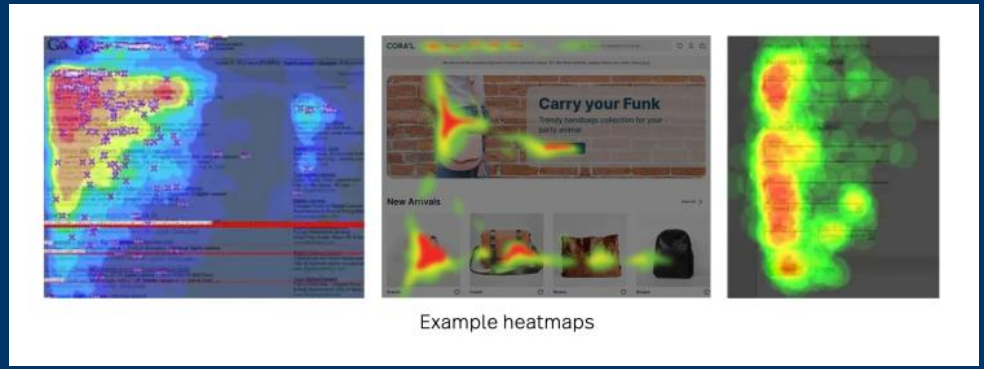
In this example, the manufacturing plant is performing well on all three metrics. They have a lower defect rate than the target and comparative values, a higher on-time delivery rate than the comparative value, and a yield rate that is very close to the target. Overall, this 3D bullet graph provides a clear and concise visual representation of the plant's performance on three different quality metrics.

Heat maps

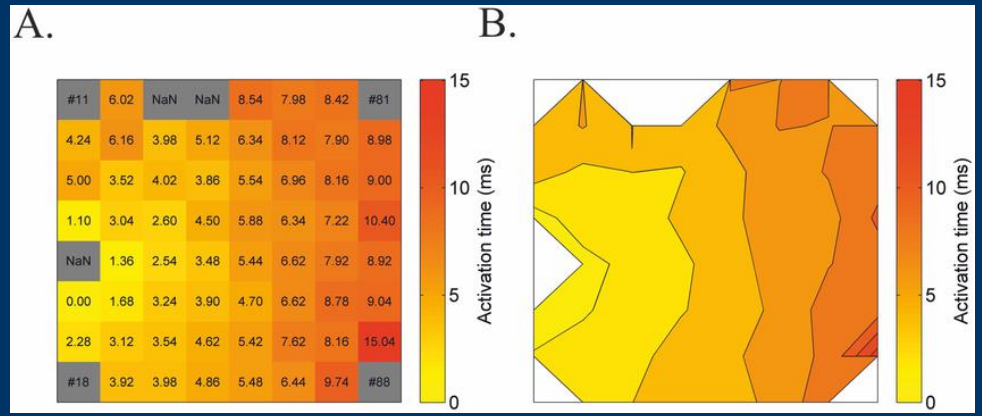
They use color coding to represent the values of a dataset, in order to display the distribution or density of data points across a geographic area, but they can also be used to show the relationship between two variables in a table or matrix format.
Each data point is represented by a colored cell or square, with the color indicating the value of the data point. Typically, a color gradient is used to represent the values, with darker or brighter colors indicating higher or lower values, respectively. The heat map can also include labels, annotations, and other features to provide more context and insights into the data.

Heat maps

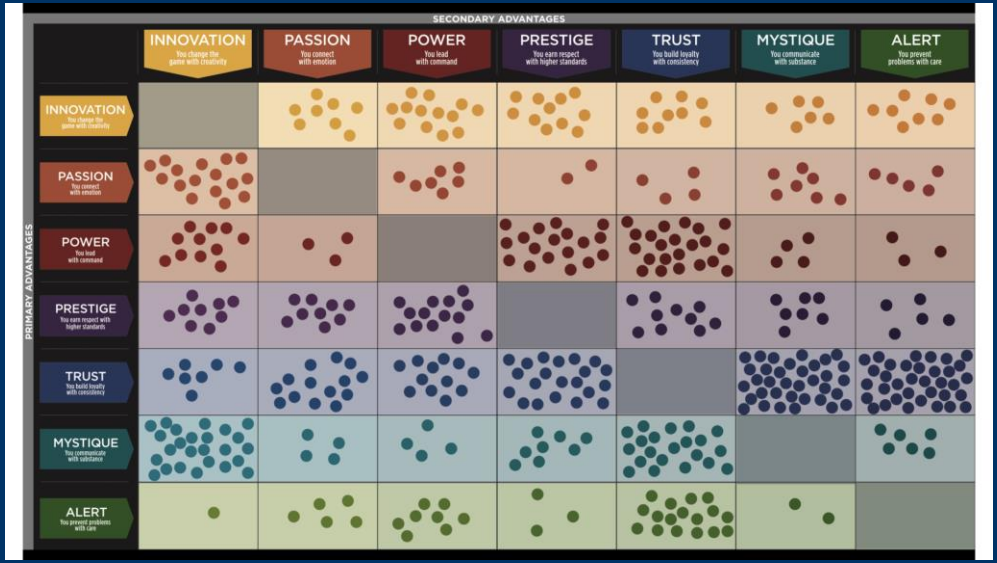


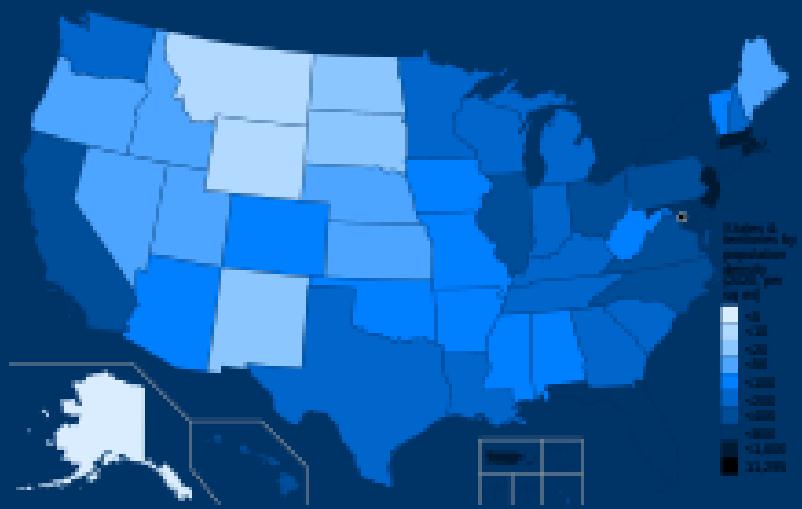


Example heatmaps



Heat maps


SECONDARY ADVANTAGES
INNOVATION
PASSION
POWER
PRESTIGE
TRUST
MYSTIQUE
ALERT
PRIMARY ADVANTAGES



Ex: Suppose we want to visualize the performance of a group of employees on several different tasks, where performance is measured on a scale of 1 to 10. We can use a heat map to display this information.

First, we need to gather the data for each employee and task. Here's a table showing the performance scores for each employee on each task:

Employee	Task 1	Task 2	Task 3	Task 4	Task 5
Employee 1	8	7	6	9	5
Employee 2	6	5	8	7	6
Employee 3	9	8	7	5	8
Employee 4	7	9	6	8	7
Employee 5	5	6	9	7	6

To create a heat map, we can use a color scale to represent the performance scores. We'll use a gradient from green to red, where green represents high performance and red represents low performance.

	Task 1	Task 2	Task 3	Task 4	Task 5
Emp 1	8 (green)	7 (green)	6 (yellow)	9 (green)	5 (red)
Emp 2	6 (yellow)	5 (red)	8 (green)	7 (green)	6 (yellow)
Emp 3	9 (green)	8 (green)	7 (green)	5 (yellow)	8 (green)
Emp 4	7 (green)	9 (green)	6 (yellow)	8 (green)	7 (green)
Emp 5	5 (red)	6 (yellow)	9 (green)	7 (green)	6 (yellow)

In this example, we can quickly see which employees are performing well and which ones are struggling. Employee 3, for example, has high scores on all tasks, while employee 5 has several low scores. The color scale makes it easy to see these patterns at a glance. Heat maps are a great way to visualize large amounts of data and identify trends and patterns in the data.

Fever Chart

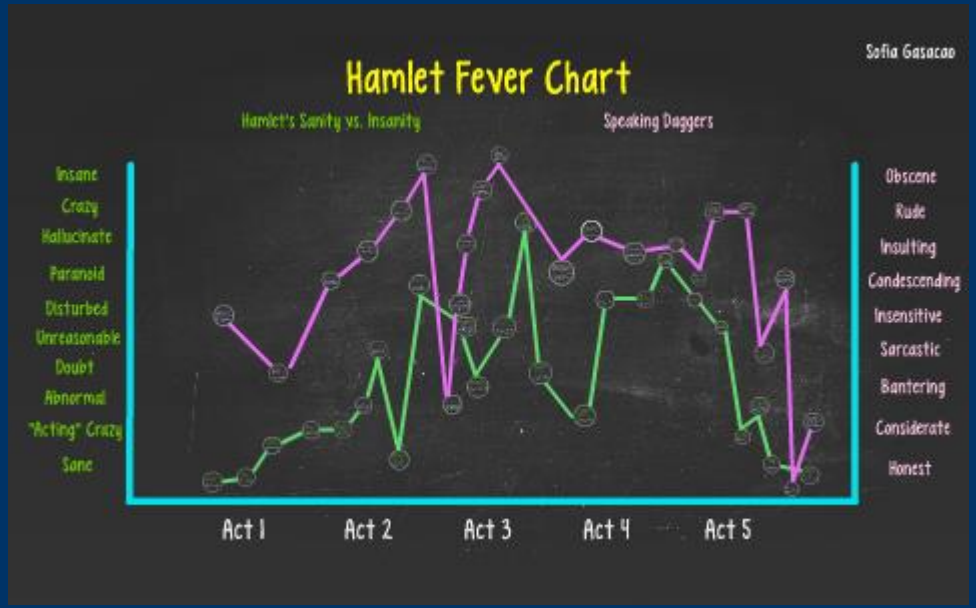
It is a graphical representation of changing data over time.

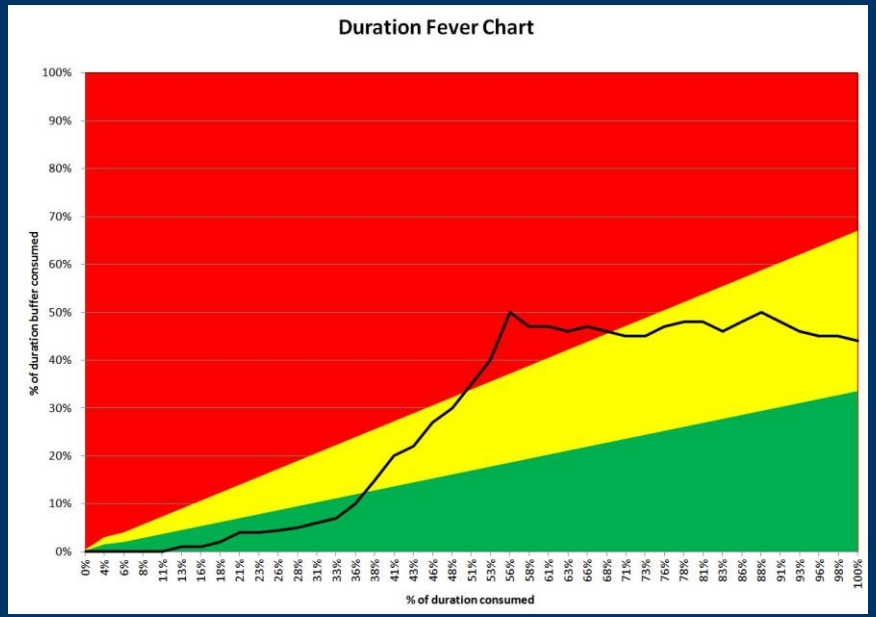
e.g. change in population in a certain region over a period of time.

When the values of variables are recorded and viewed over a long period of time, it is difficult to derive patterns or trends from plain data.

However, when the same data is represented in a fever chart, it becomes easier to spot trends or patterns

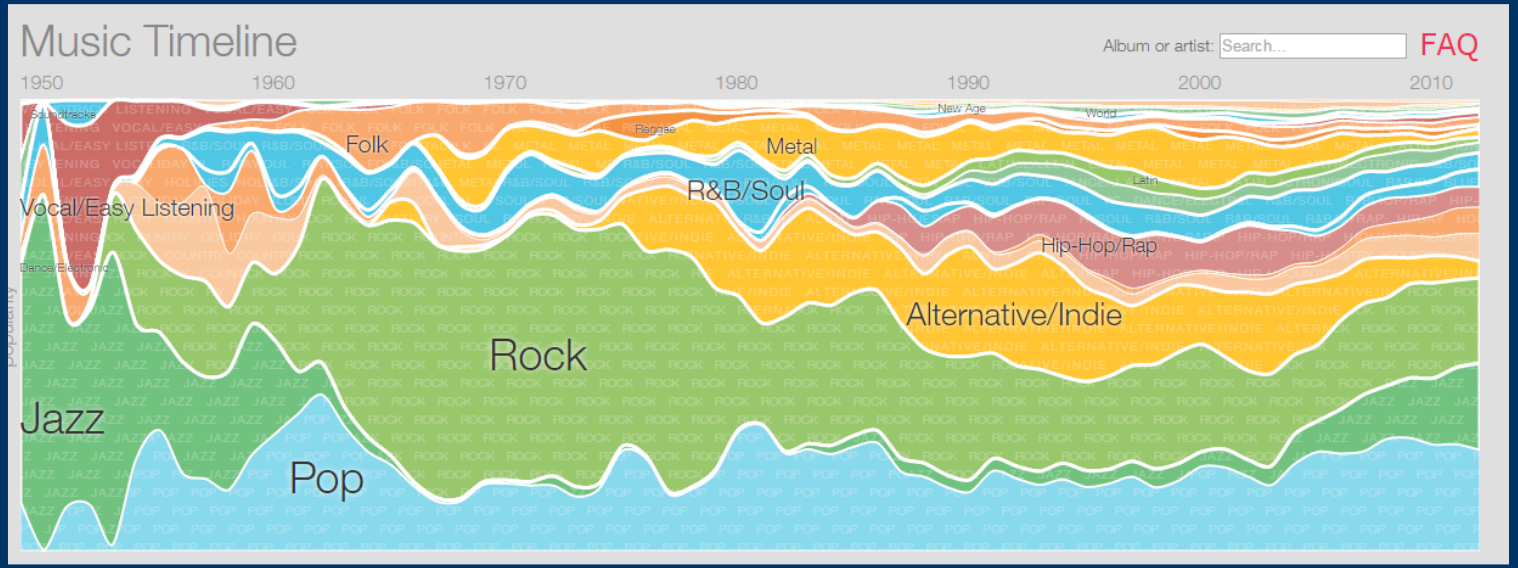
Fever charts





Fever Charts

Variation of popular music from 1950 - 2010



Ex: Suppose we want to visualize the body temperature of a patient over the course of a week. We can use a fever chart to display this information.

First, we need to gather the data for the patient's body temperature at different times throughout the week. Here's a table showing the patient's temperature readings:

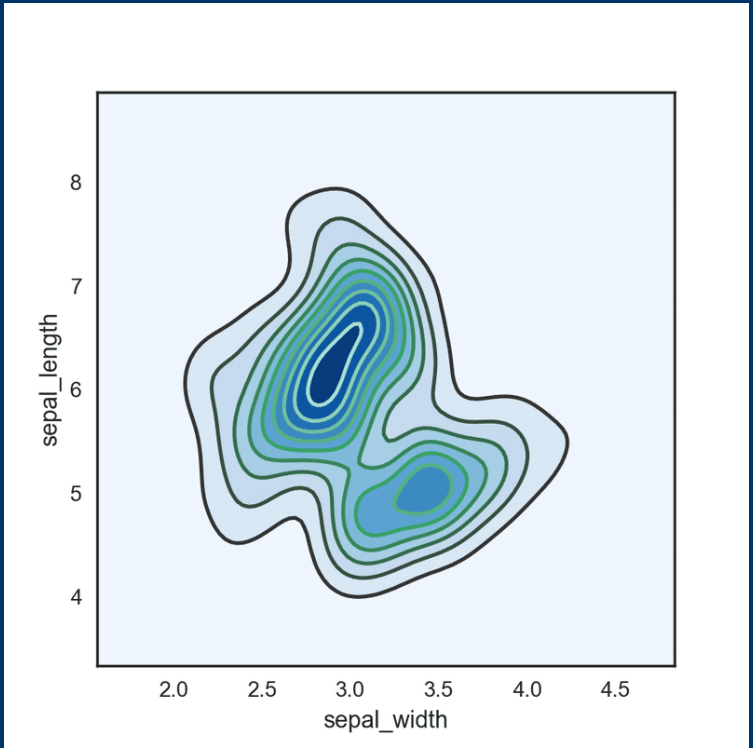
To create a fever chart, we can plot the temperature readings on a graph over time, with the temperature on the y-axis and time on the x-axis. We can use different colors to indicate different temperature ranges, such as green for normal temperature, yellow for elevated temperature, and red for high fever.

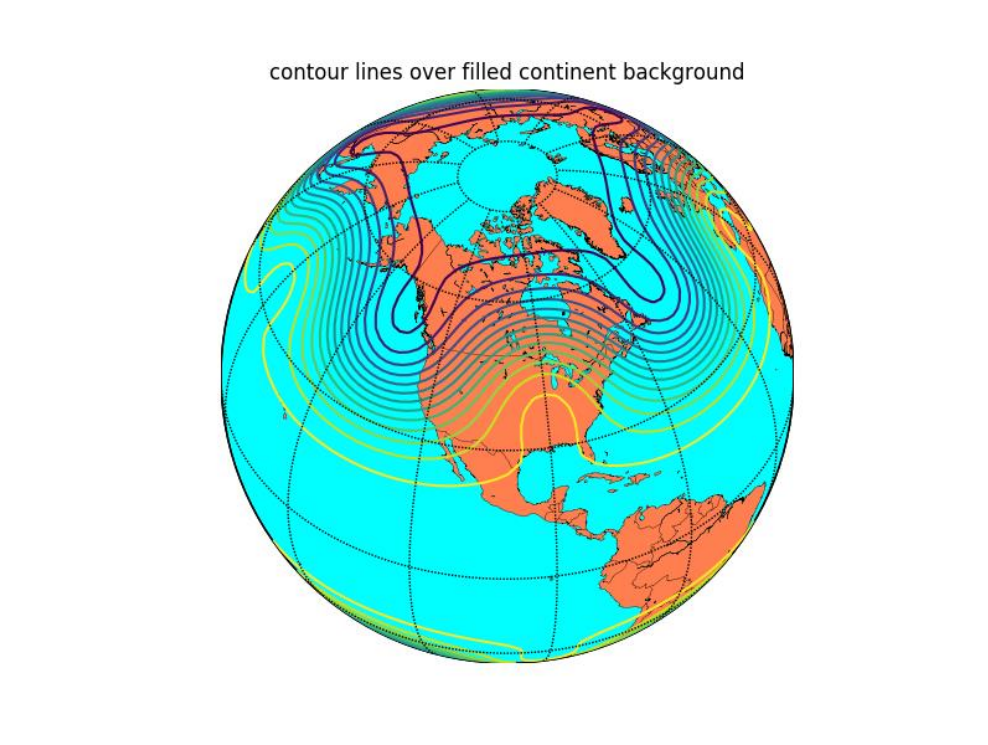
Day	Time	Temperature (°F)
Monday	8:00 AM	98.7
	12:00 PM	99.1
	4:00 PM	99.5
	8:00 PM	100.2
Tuesday	8:00 AM	99.0
	12:00 PM	99.3
	4:00 PM	99.7
	8:00 PM	100.4
Wednesday	8:00 AM	99.4
	12:00 PM	99.7
	4:00 PM	100.2
	8:00 PM	101.1
Thursday	8:00 AM	99.9
	12:00 PM	100.3
	4:00 PM	101.2
	8:00 PM	101.5
Friday	8:00 AM	100.1
	12:00 PM	100.5
	4:00 PM	101.0
	8:00 PM	101.3
Saturday	8:00 AM	99.5
	12:00 PM	99.8
	4:00 PM	100.1
	8:00 PM	100.5
Sunday	8:00 AM	98.9
	12:00 PM	99.2
	4:00 PM	99.7
	8:00 PM	100.1

2D density plots

Also known as density contour plots, show the density of data points in a two-dimensional space. They are similar to heat maps in that they use color to indicate the intensity or concentration of data points in different areas of the plot. However, density plots also include contour lines that connect areas of equal density, allowing for a more detailed representation of the distribution of data, each data point is assigned a weight or probability density, which is used to estimate the density of points in a given area of the plot. The plot is typically divided into a grid of smaller cells or bins, and the density of data points in each cell is calculated and represented using color and contour lines.

2D density plots





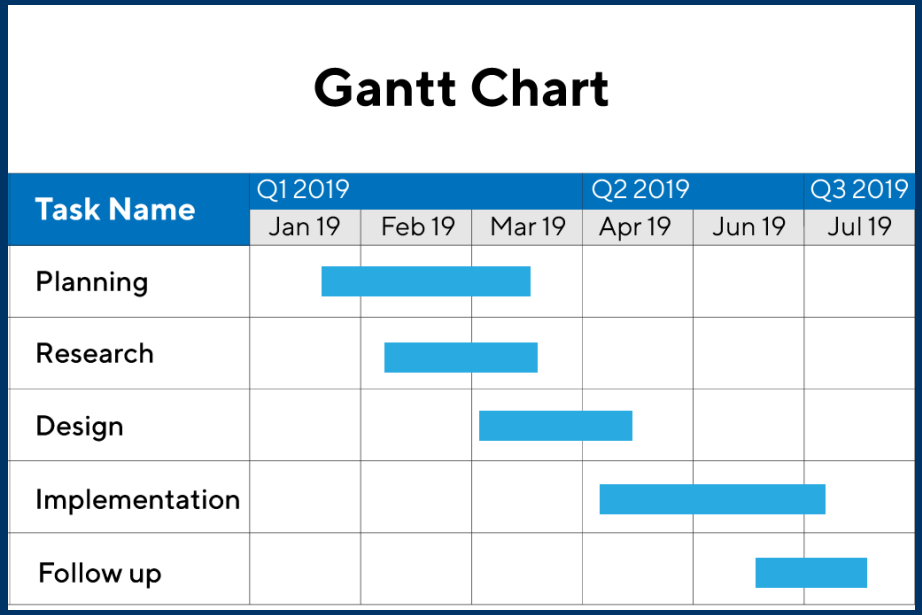
Gantt charts

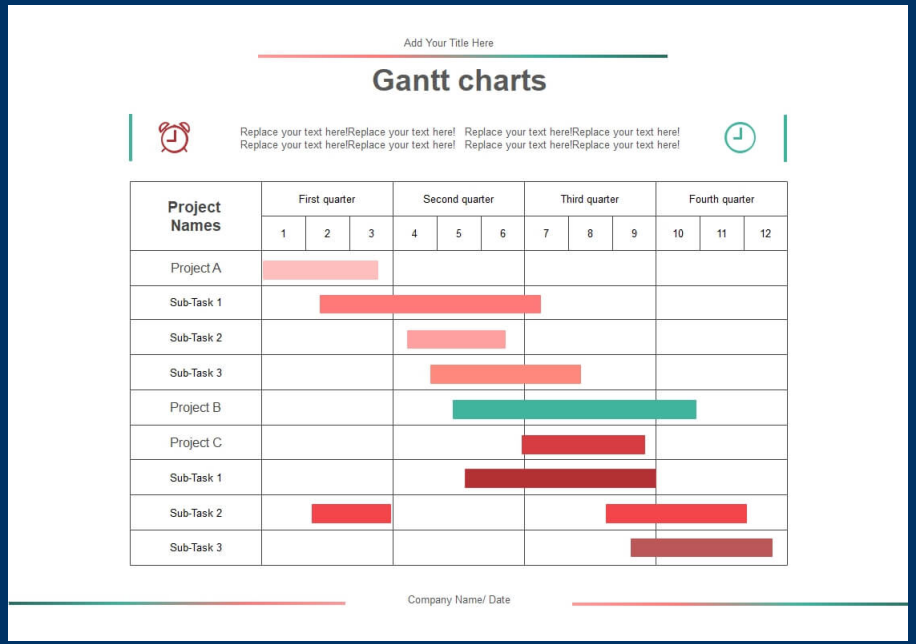
Named after Henry Gantt, it is used in project management to visualize the timing and duration of tasks or activities within a project.

Each task or activity is represented by a horizontal bar that spans the duration of the task. The bars are typically arranged in rows or columns to represent different categories or phases of the project, such as project milestones, resources, or departments. It may also include task dependencies, start and end dates, and progress status.

Useful in project planning and scheduling, as they allow project managers to visualize the timing and dependencies of tasks and to allocate resources more effectively.

Gantt charts





Ex: Suppose you are managing a software development project with the following tasks and durations:

Task A: Design the user interface (5 days)

Task B: Code the back-end logic (10 days)

Task C: Write unit tests (3 days)

Task D: Perform integration testing (7 days)

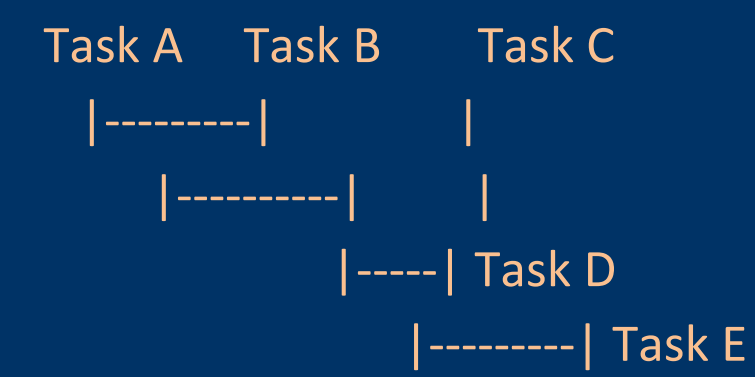
Task E: Fix bugs and issues (5 days)

To create a Gantt chart for this project, you would first list the tasks in order and create a horizontal timeline that spans the entire project duration. For simplicity, let's assume that the project will take 30 days to complete:

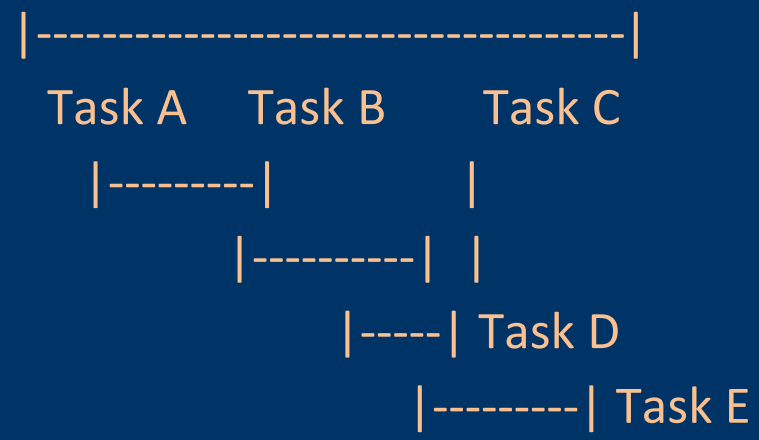
|----------|

Task A Task B Task C Task D Task E

Next, you would mark the start and end dates for each task along the timeline. For example:



Finally, you would add dependencies between tasks by linking them with arrows. For example, Task B cannot start until Task A is completed, and Task D cannot start until both Task B and Task C are completed. The resulting Gantt chart would look like this:

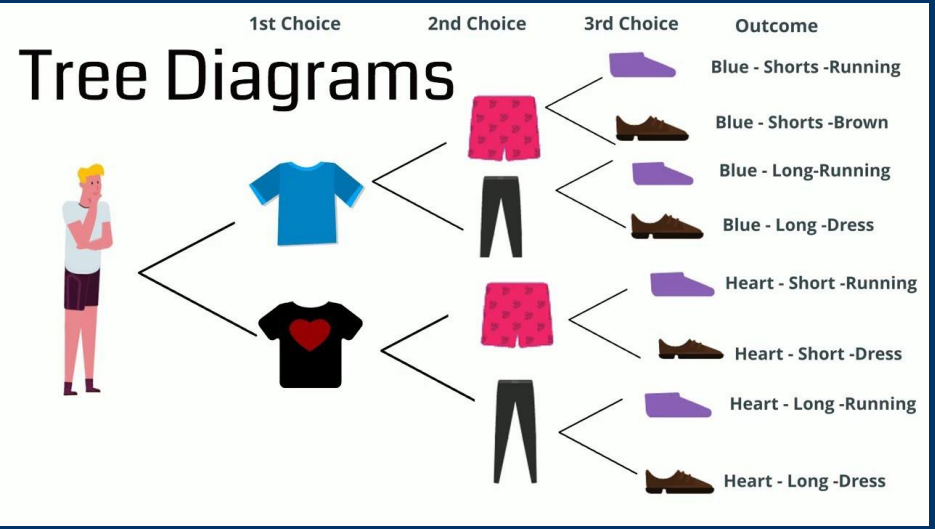


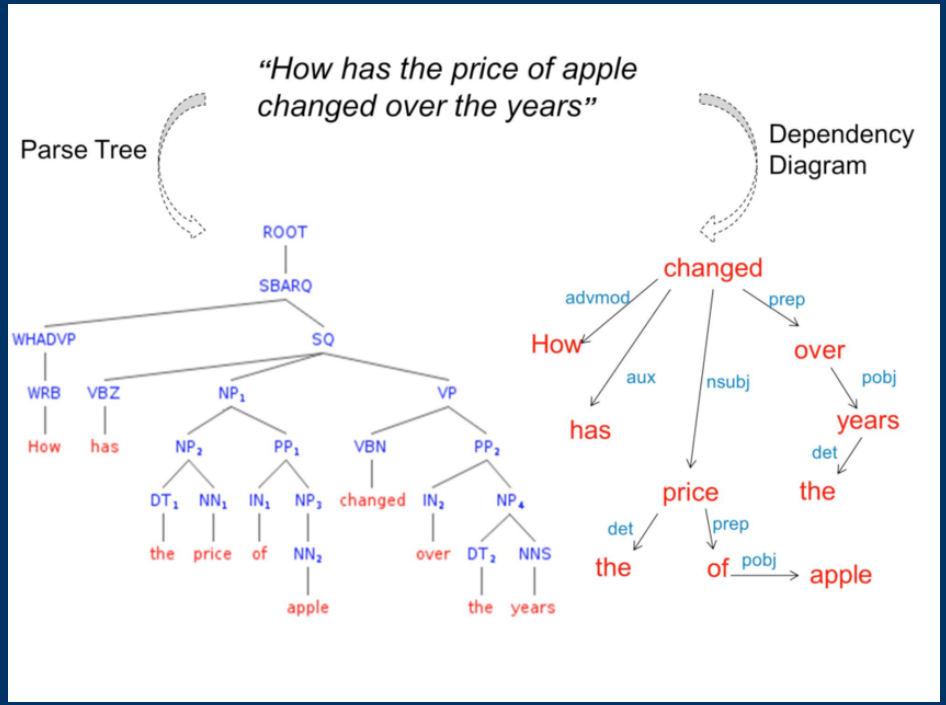
This Gantt chart shows the start and end dates for each task, as well as the dependencies between tasks. It can help you visualize the project schedule and identify potential scheduling conflicts or delays.

Tree Diagrams

It represents hierarchical relationships or structures. It consists of a series of nodes or boxes that are connected by lines or branches to show the relationships between them. Each node represents a category, attribute, or subcategory, and the branches represent the connections or relationships between them. It has a root node at the top that represents the main category or concept, and subsequent nodes that represent subcategories or attributes that are related to the main category. The diagram can be expanded or collapsed to show more or less detail, depending on the level of information needed.

Tree Diagrams





Ex: Suppose you are analyzing the potential outcomes of a decision to launch a new product. You identify three possible outcomes: success, moderate success, or failure. For each outcome, you identify three possible factors that could influence the result. For simplicity, let's assume that the factors are independent of each other:

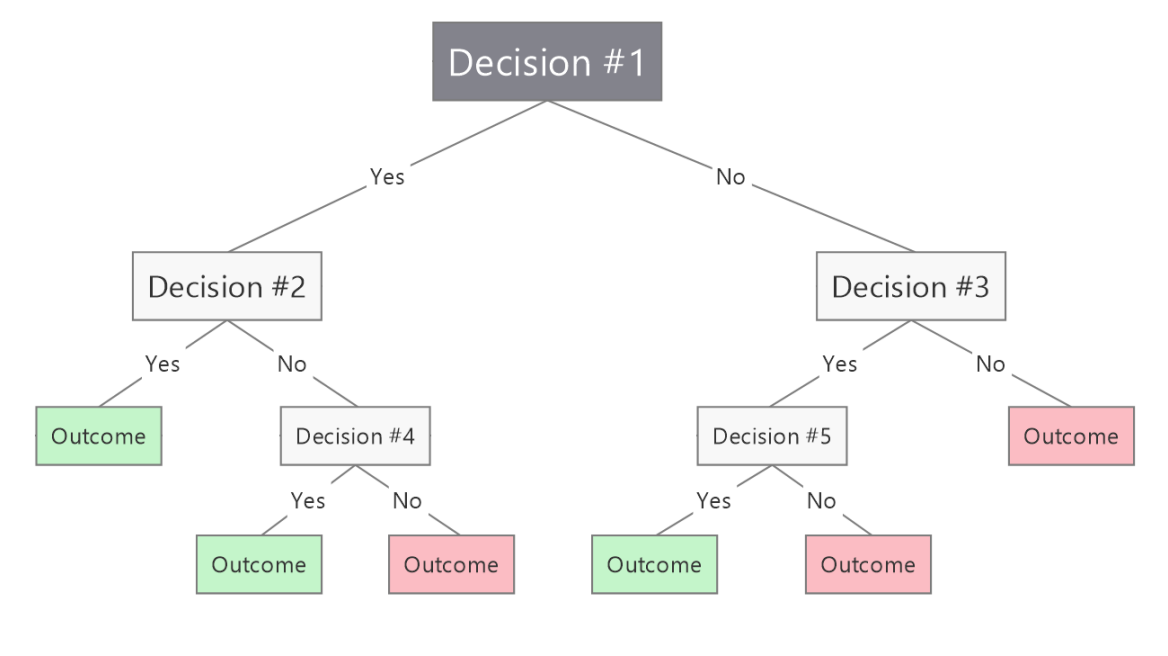
Success: high demand, positive customer reviews, effective marketing

Moderate success: moderate demand, mixed customer reviews, moderate marketing

Failure: low demand, negative customer reviews, ineffective marketing

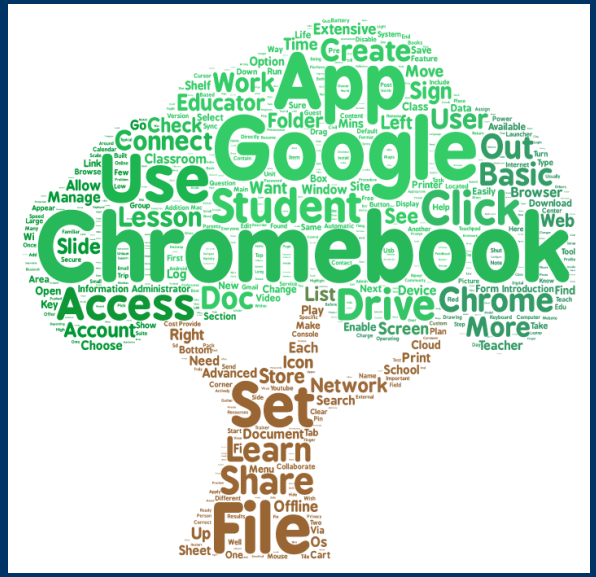
Create a tree diagram for this decision. Start by drawing a single node representing the decision to launch the product

Decision Trees



Word Clouds

It is a collection, or cluster, of words depicted in different sizes.



Ex: Suppose you want to create a word cloud to visualize the most common words in a customer feedback survey. You have collected feedback from 50 customers and compiled a list of their responses, which includes both positive and negative feedback. For simplicity, let's assume that you have already cleaned and pre-processed the text data.

Here are the top 10 most common words in the feedback survey and their respective frequencies:

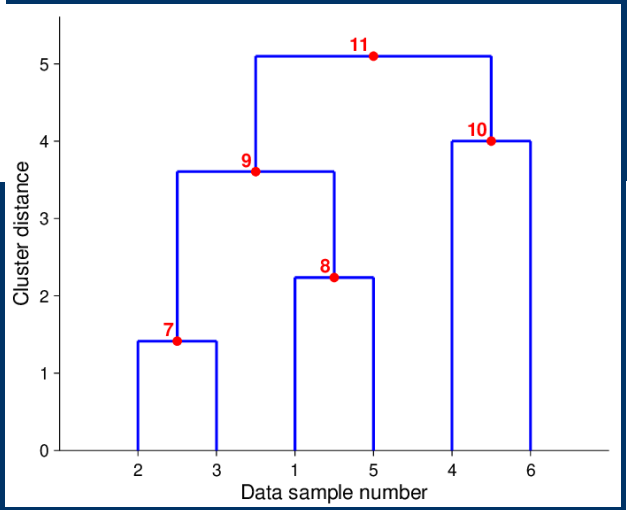
Excellent (20)
Service (18)
Product (16)
Quality (14)
Good (12)
Poor (10)
Satisfaction (8)
Price (6)
Recommend (4)
Improvement (2)

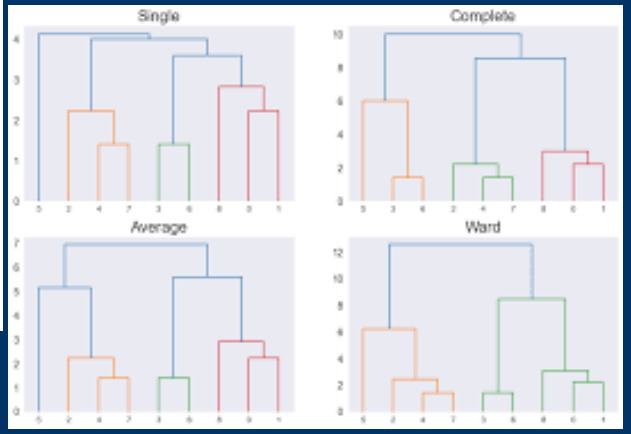
To create a word cloud for this data, you would first choose a visualization tool that allows you to create word clouds (e.g. WordCloud in Python or Word Cloud Generator in Microsoft Excel). Then, you would input the top 10 words and their frequencies into the tool and adjust the font, color, and layout settings as desired.

Dendrograms

It is tree diagram used to represent the hierarchical clustering of objects or data points based on their similarities or dissimilarities. It consists of a series of branching clusters or nodes that represent groups of objects or data points that are more similar to each other than to objects or data points in other clusters. Each object or data point is represented by a leaf node at the bottom of the diagram, and the clusters are formed by merging the most similar objects or data points together into larger and larger groups. The height of each node represents the level of similarity between the clusters, with the tallest nodes indicating the largest differences between clusters.

Dendrogram


11
10
9
8
7
Cluster distance
Data sample number
0
1
2
3
4
5
2
3
1
5
4
6


Single
Complete
Average
Ward

Matrix charts

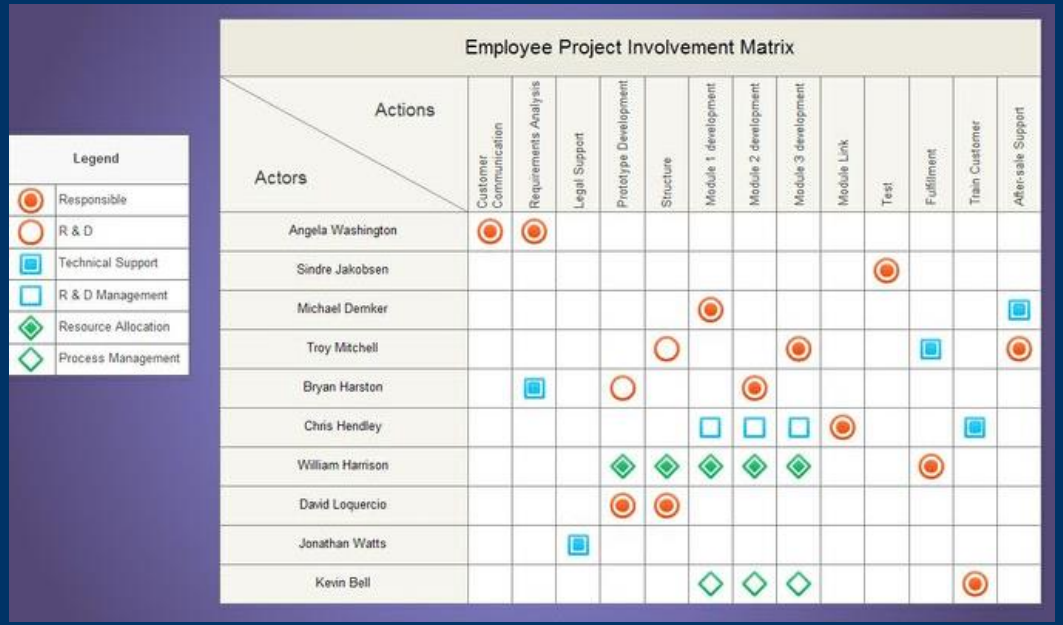
They consist of a grid with two or more dimensions. Each axis represents a different variable or category, and the intersections between the axes show the relationships between them. The cells within the matrix can be color-coded, shaded, or filled with symbols to represent the strength or importance of the relationship between the variables.

There are different types of matrix charts, including the SWOT (Strengths, Weaknesses, Opportunities, Threats) matrix, the BCG (Boston Consulting Group) matrix, and the Eisenhower matrix. These charts can be used for a variety of purposes, such as strategic planning, risk management, and prioritization.

Matrix Chart

FMEA	◬	◎	O	◬	◬
SPC	O	O	◬	◎	
DMADV		◬	◎	◬	◬
DOE	◎	◬		O	◎
Lean Principles	O		◬	◬	O
Employee	**Tom**	**Jack**	**Bob**	**Matt**	**Paul**
Project 1	◎		◬	◬	◎
Project 2		O		◬	◬
Project 3	◬	◎	O		◎
Project 4			◎	O	
Project 5	◎	◬	◎		O

Symbol	◎	O	◬
Value	9	3	1
Relationship	Strong	Medium	Weak



Ex: Suppose you are analyzing the performance of a sales team and you want to identify which salespeople are performing well and which ones need additional training or support. You have collected data on the number of sales made by each salesperson and the average revenue per sale. For simplicity, let's assume that you have 5 salespeople on your team:

Salesperson A: 10 sales, $100 average revenue per sale

Salesperson B: 8 sales, $120 average revenue per sale

Salesperson C: 12 sales, $80 average revenue per sale

Salesperson D: 6 sales, $150 average revenue per sale

Salesperson E: 9 sales, $90 average revenue per sale

To create a matrix chart for this data, you would start by creating a 2x2 grid with one axis representing the number of sales and the other axis representing the average revenue per sale. You would then plot each salesperson's data as a point in the chart, with the x-coordinate representing the number of sales and the y-coordinate representing the average revenue per sale.

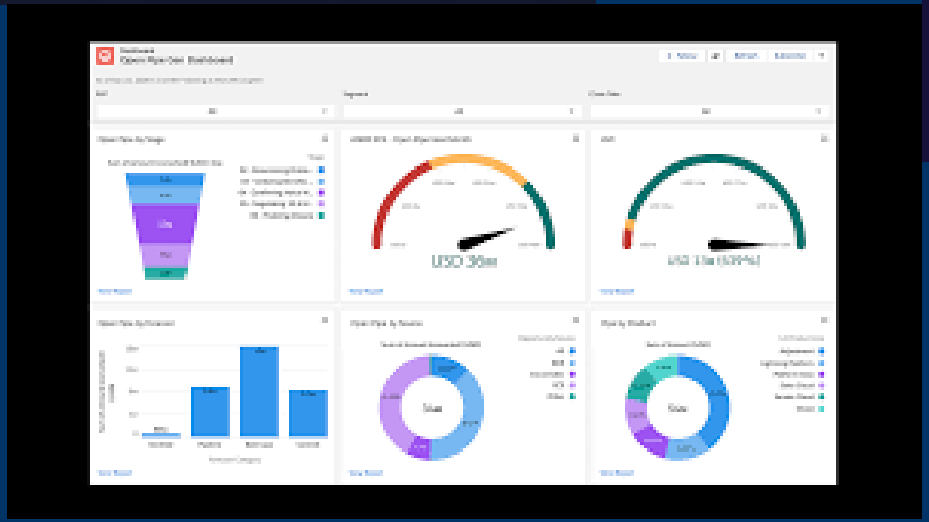
This matrix chart shows the number of sales and the average revenue per sale for each salesperson, and allows you to quickly identify the best and worst performers. In this example, Salesperson B has the highest average revenue per sale and Salesperson C has the highest number of sales, while Salesperson D has the lowest number of sales and the highest average revenue per sale. You can use this information to provide targeted feedback and support to each salesperson based on their individual strengths and weaknesses.

Dashboards

They provide a real-time snapshot of key performance indicators (KPIs) and other important metrics used in business and organizational contexts to monitor progress towards goals, track trends over time, and identify areas for improvement.

They consist of a series of charts, graphs, tables, and other visualizations that show data from different sources in a single, unified view. They can be customized to display the most relevant information for a particular user or audience, and can be updated automatically as new data becomes available.

Dashboards




Sales
£297K
this month
£16K vs last month
$9.6K
today
$20.6K
yesterday
Biggest deals this month
Alice $8,600
Jared $8,500
Heather $7,540
Shaun $7,450
Marsha $6,530
Jared $4,565
Heather $4,560
Polly $4,215
Dalisu $3,560
Social followers
19.5K
LinkedIn
11 v yday
10.5K
Twitter
22 v yday
Website (past 7 days)
27.2K
Users
1.6K vs last week
126
Enquiries
28 vs last week
NPS (past 30 days)
0
100
61
Recent feedback
ok
Very Helpfull
very good "thumbs up"
Twitter Mentions
To celebrate #FoM22, we're excited to offer 25% off expert speaker, Neil Perkin's latest release, Agile...
Festival of Marketing - a day ago
RT @FestofMarketing: Really inspiring session at #FoM22 with @luciecave and Munroe Bergdorf, speaking on...
Bauer Media Group UK - 3 days ago
RT @TweetMRSpride: We're at @FestofMarketing today and looking forward to the opening address...
Noreena Hertz - 3 days ago
Current US visitors
Daily pulse dashboard
Powered by Geckoboard
14:43

Ex: Suppose you are managing a social media marketing campaign for a new product launch and you want to track the campaign's performance in real-time. You have collected data on various metrics such as the number of clicks, impressions, engagement rate, and conversion rate. For simplicity, let's assume that you have collected data for the past week and you want to visualize it on a dashboard.

To create a dashboard for this data, you would first choose a dashboard tool (e.g. Tableau, Power BI, or Google Data Studio) and create a new dashboard. Then, you would add various charts and visualizations to the dashboard that show the key metrics of the campaign.

This dashboard should include several charts and visualizations that show the performance of the social media marketing campaign:

A line chart that shows the number of clicks and impressions over time

A pie chart that shows the distribution of engagement by social media platform

A bar chart that shows the conversion rate by age group

A table that shows the top-performing posts by engagement rate

Each chart and visualization can be customized to show the data in different ways (e.g. by date range, by social media platform, by geographic region, etc.). The dashboard allows you to quickly and easily track the performance of the campaign and make data-driven decisions to optimize your marketing strategy.

Thank You
and
Happy Visualization
?